EDITED VOLUMES

Theo Hug, Günther Pallaver (Hg.)

Talk with the Bots

Gesprächsroboter und Social Bots im Diskurs



innsbruck university press

EDITED VOLUME SERIES



innsbruck university press

Theo Hug, Günther Pallaver (Hg.)

Talk with the Bots

Gesprächsroboter und Social Bots im Diskurs

Theo Hug
Institut für Medien, Gesellschaft und Kommunikation, Universität Innsbruck
Sprecher des interfakultären Forums *Innsbruck Media Studies* an der Universität Innsbruck

Günther Pallaver
Institut für Politikwissenschaft, Universität Innsbruck
Leiter des Instituts für Medien, Gesellschaft und Kommunikation, Universität Innsbruck

Gedruckt mit finanzieller Unterstützung der Moser Holding AG, der Austria Presse Agentur (APA), des Interfakultären Forums Innsbruck Media Studies sowie des Vizerektorats für Forschung der Universität Innsbruck.







Universität Innsbruck
1. Auflage

www.uibk.ac.at/iup
ISBN 978-3-903187-29-0

Inhaltsverzeichnis

Editorial

Theo Hug, Günther Pallaver

Immer öfter kommuniziert Software mit und für uns. Bots übernehmen oft unbemerkt komplexe Aufgaben und virtuelle Assistenzen verarbeiten die gesprochene Sprache. Die Technologieentwicklung gestattet uns mittlerweile, die Dienste von digitalen, oftmals sprachbasierten Assistenzsystemen in unseren Alltag zu integrieren. Besonders im Bereich der politischen Propaganda und Public Relations sind Social Bots seit einiger Zeit zu zweifelhafter Berühmtheit gelangt. Eingesetzt als Instrument zur Meinungsmache verbreiten sie automatisiert Nachrichten in sogenannten „sozialen" Netzwerken. Der interdisziplinär ausgerichtete Band befasst sich mit dieser Thematik und sucht nach Antworten auf Fragen wie die folgenden: Wie können wir die Interaktion mit künstlichen Intelligenzen von jener mit Menschen unterscheiden? Inwieweit ist das wichtig? Was bedeutet diese Entwicklung für das Alltagsleben und für den Bereich der Kommunikation? Welche Bedeutung hat der verbreitete Einsatz von Bots als neue Schnittstelle zwischen Mensch und Maschine?

Diese und ähnliche Fragestellungen standen im Wintersemester 2017/18 im Zentrum des Medientags 2017 und der Ringvorlesung mit Übungscharakter „Mit Computer kommunizieren: Gesprächsroboter und Social Bots im Diskurs." Sowohl der Medientag zum Thema „Talk with the Bots. Die Maschine versteht dich" als auch die Ringvorlesung wurden vom interfakultären Forschungsforum *Innsbruck Media Studies* in Kooperation mit unseren Veranstaltungspartnern, der Moser Holding AG und der Austria Presse Agentur an der Universität Innsbruck durchgeführt.

Wenn hier von „Bots" die Rede ist, dann ist damit nicht eine saloppe Ausdrucksform für alle Arten von Robotern gemeint – von den großen Industrierobotern über Mähroboter für den Garten, Staubsaugerroboter und Kochroboter bis zu den programmierbaren Nanomaschinen, für die Jean-Pierre Sauvage, Fraser Stoddart und Bernard Feringa 2016 den Chemie-Nobelpreis erhalten haben. Im Zentrum der Überlegungen stehen hier *Social Bots* – also jene Programme, die in „sozialen" Netzwerken menschliche Nutzer simulieren – und *Chatbots* als text- und sprachbasierte Dialogsysteme, die häufig in Verbindung mit Avataren benutzt werden.

Chatbots gibt es durchaus schon länger – einer der ersten Chatbots war ELIZA (Joseph Weizenbaum, 1966) – inspiriert von psychotherapeutischem Kommunikationsgeschehen. Wer in den 60er und 70erJahren an der Universität Innsbruck Psychologie studiert hatte, konnte eine analoge Version bei Prof. Ivo Kohler ausprobieren.

Solche Systeme werden heute auch als virtuelle persönliche Assistenten bezeichnet – mancherorts wird auch von „virtuellen Butlern" oder „digitalen Sklaven" gesprochen. Die Konnotationen solcher Ausdrucksformen machen schnell deutlich, dass hier medien-, maschinen-, informations- und technikethische Fragen eine große Rolle spielen.

In der Softwareindustrie ist allerdings seit einiger Zeit von einem Paradigmenwechsel die Rede, von „Messaging“ als neuer Plattform und „Bots“ als den neuen Apps, von neuen Entwicklungsumgebungen und Geschäftsmodellen, sowie von neuen Nutzungsmöglichkeiten für EndanwenderInnen und neuen Werbestrategien. Es geht bei der Themenstellung des Sammelbandes jedoch nicht in erster Linie um die Erweiterung von Handlungsspielräumen für Unternehmen im digitalen Kapitalismus und deren Grenzen – es geht um viele verschiedene Lebensbereiche sowie Ebenen der Modellierung und Reflexion und nicht zuletzt um wissenschaftliche Sondierungen in einem Forschungsbereich, der erst in den Anfängen steckt.

Der vorliegende Band führt theoretische und anwendungsorientierte Beiträge dieser beiden Veranstaltungen zusammen, ergänzt durch einen abschließenden Beitrag von Claudia Paganini aus dem Kreis des Forums *Innsbruck Media Studies.*

Nach den Grußworten von *Tilmann Märk*, Rektor der Leopold-Franzens-Universität Innsbruck, *Hermann Petz*, CEO der Moser Holding und *Clemens Pig,* CEO der APA, eröffnet *Timo Kaerlein* den Band mit dem Beitrag *Social Bots und die Formalisierung von Sozialität auf Online-Plattformen.* Dass das Verhalten von Bots als soziale Dimension wahrgenommen werden kann, ist primär Ausdruck eines Prozesses der Formalisierung von Sozialität. Der Beitrag führt diesen Prozess auf die Urszene der Mensch-Maschine-Kommunikation zurück, auf den erstmals 1950 vorgeschlagenen Turingtest, der als Gradmesser der Intelligenz eines Computerprogramms dienen sollte.

Um linguistische Perspektiven Z*ur sprachlichen Interaktion mit Chatbots* geht es im Beitrag von *Netaya Lotze.* Sie diskutiert dabei eine völlig neue kommunikative Situation, da artifizielle Systeme mit natürlichsprachlichem Interface den Anschein menschlicher KommunikationspartnerInnen erwecken.

Der Maschinen- und Informationsethiker *Oliver Bendel* geht in seinem Beitrag *Chatbots als Artefakte der Maschinenethik* der Frage nach, ob und wie Maschinen ethisch agieren können. Dabei legt er dar, welche Potenziale und welche Vor- bzw. Nachteile die Umsetzung dieser Artefakte in der Maschinenethik und darüber hinaus hat.

Der Beitrag von *Thorsten Büchner* und *Imke Troltenier* zu *Sprachassistenzsysteme. Expertise Blinder und Sehbehinderter* stellt exemplarisch dar, wie Menschen mit Blindheit und Sehbehinderung Informationen aufnehmen und in welcher Art und Weise Sprachassistenz dabei auf vielfältige Weise unterstützen kann.

Andrea Knaut setzt sich mit der Frage *Können Künstliche Neuronale Netzwerke denken?* auseinander. Künstliche Neuronale Netzwerke sind dabei nur eines von verschiedenen Modellparadigmen, auf das die Computerlinguistik und die Informatik-Disziplin Künstliche Intelligenz zurückgreifen. Inzwischen werden in bestimmten Kontexten auch im sprachlichen Bereich menschliche Fähigkeiten übertroffen oder ergänzen sie. Aber denken sie damit schon?

Rainer Leschkes weist in seinem Beitrag *Subjektlose Verantwortung – Zur Ethik autonomer Systeme* darauf hin, dass sowohl Maschinen als auch Software nicht selten die Eigenschaften von Subjekten zugebilligt werden. Damit würden sie über all die Eigenschaften verfügen, die die Subjektphilosophie seit der europäischen Aufklärung dem Menschen zuzubilligen bereit war.

Den Band schließt der Beitrag *Können Chatbots lügen?* von *Claudia Paganini.* Um eine Antwort auf diese Frage zu finden, werden verschiedene Merkmale der Lüge herausgearbeitet: Dabei zeigt sich trotz einzelner Einschränkungen, dass es durchaus angemessen ist, von lügenden Chatbots zu sprechen.

Sowohl die Veranstaltungsorganisation als auch die Herausgabe des Sammelwerkes waren ein kollaboratives Unterfangen, für das wir in mehrfacher Hinsicht zu danken haben: Ganz besonders danken wir unseren beiden Veranstaltungspartnern, der Moser Holding AG – insbesondere Herrn Mag. Hermann Petz, und der Austria Presse Agentur, Herrn Dr. Clemens Pig, für ihre Unterstützung, und Frau Mag. Lisa Berger-Rudisch, Frau Barbara Rauchwarter, Herrn Patricio Hetfleisch und Herrn Mag. (FH) Norbert Adlassnigg für die sehr anregende und konstruktive Zusammenarbeit, Frau Ulrike Pfeiffenberger und Herrn Marco Hermann für die umsichtige Organisation sowie allen Helfern und Helferinnen am Ort für die technische Betreuung, Herrn Georg Laich vom ORF für die Moderation der Podiumsdiskussion, Frau Martina Posch und Frau Jennifer Wachter für die studentische Mitarbeit bei der Ring-Vorlesung sowie den MitarbeiterInnen vom Büro für Öffentlichkeitsarbeit und Kulturservice für die gute Zusammenarbeit. Die Publikation wäre ohne die finanzielle Unterstützung des Vizerektorats für Forschung der Universität Innsbruck sowie der beiden Veranstaltungspartner nicht möglich gewesen. Zu danken haben wir weiters Thomas Walli für die Layout-Unterstützung sowie Dr. Birgit Holzner und Carmen Drolshagen von *innsbruck university press* für die verlegerische Betreuung.

Innsbruck, im Juli 2018 Theo Hug & Günther Pallaver

Grußworte zum Medientag 2017

Tilmann Märk, Rektor der Universität Innsbruck

Sehr geehrte Damen und Herren,
liebe KollegInnen,

Kommunikation ist eines der großen Themen unserer Zeit. Nicht nur im Alltag, sondern auch an der Universität.

Exzellente Forschung und Lehre, Hauptaufgaben der Universität, brauchen Zeit, Geld und Köpfe. Aber was nützen die interessantesten Ergebnisse, wenn sie nicht entsprechend kommuniziert werden? Eine wichtige Möglichkeit der Kommunikation, neben der Veröffentlichung als Buch oder in Zeitschriften ist die Präsentation und kritische Diskussion auf Tagungen. QED – quod erat demonstrandum.

Aber mittlerweile kommunizieren wir fast dauernd und jederzeit sowohl privat als auch beruflich. Face to Face oder über die uns ständig begleitenden Medien. Um diesem Tempo gerecht zu werden, verändern sich auch unsere Gewohnheiten und die notwendigen Technologien bzw. umgekehrt haben die letzteren dies erst ermöglicht und getriggert. Man könnte in diesem Zusammenhang formulieren: Die Welt wandelt sich und wir uns mit ihr.

Neue Technologien begegnen uns in der Wissenschaft und später auch im Alltag. Viele Vorarbeiten, die auch an der Universität geleistet werden, finden sich später in einem Produkt am Markt wieder. Alexa, Siri und Co. sind längst keine Unbekannten mehr und in vielen Haushalten zu finden. Die Arbeit an bzw. mit komplexen Technologien bereichert, erleichtert oder verändert unseren Alltag. Sogenannte Bots übernehmen unbemerkt immer komplexere Aufgaben für uns als NutzerInnen. Sprachsteuerungen werden auch im Sinne des „Smart Living“ Einzug in unser Leben finden. Alexa wird uns helfen die Kochuhr zu programmieren, das Licht im Wohnzimmer zu dimmen oder den Einkaufszettel zu schreiben.

D. h., wir sind umgeben von einer sich ständig verändernden Welt und können noch nicht ermessen, wohin diese Reise gehen wird. Medien und Produkte werden stetig weiterentwickelt und es bleibt spannend, welche Neuerungen uns noch begegnen werden. Auch die Wissenschaft leistet hier natürlich wesentliche Arbeiten zur ständigen Entwicklung. Die Kommunikation und die sich eröffnenden Möglichkeiten werden sich weiterhin rasant entwickeln und auch die Schnittstelle zwischen Mensch und Maschine wird eine noch zentralere Rolle bekommen.

Insofern ist es durchaus konsequent, dass die LFUI im neuen Entwicklungsplan diesem Thema unter dem Stichwort Digital Sciences breiten Raum eingeräumt hat. Wir haben

bereits jetzt hohe Kompetenz in Teilbereichen und haben zusätzlich vor, in ein paar Wochen bis zu zehn Nachwuchsprofessuren auf diesem Gebiet auszuschreiben und hochkarätig zu besetzen. Dabei wird es um Themen gehen wie Datenspeicherung, Datenanalyse, Datenanwendung, Künstliche Intelligenz, Maschinenlernen, Robotics, Digital Humanities und Digital Transformations.

In diesem Zusammenhang möchte ich nicht unerwähnt lassen, dass die LFUI auf diese Entwicklung auch insofern reagiert hat, dass wir im Lehramtsstudium eine Spezialisierungsmöglichkeit im Bereich Medienpädagogik und ein neues Institut für Medien, Gesellschaft und Kommunikation an der Fakultät für Soziale und Politische Wissenschaften eingerichtet haben, die heute mitveranstaltenden Kollegen Theo Hug und Günther Pallaver sind Gründungsprofessoren dieses Instituts.

Abschließend halte ich fest: Ich gehe daher davon aus, dass sich der Medientag 2017 im Rahmen der Präsentationen und Diskussionen im Detail und sowohl aus wissenschaftlicher als auch aus praktischer Sicht mit den Fragen (i) inwieweit sich diese Veränderungen auf uns selbst auswirken werden, (ii) ob es eine Erleichterung oder eine Bürde ist, (iii) welche Vorteile sie für uns bringen und (iv) vor allem wie diese kritisch zu hinterfragen sind, auseinandersetzen wird. Dafür viel Erfolg.

Grußworte zur Eröffnung des Medientags 2017

Hermann Petz, Vorstandsvorsitzender der Moser Holding AG

Sehr geehrter Herr Rektor,

sehr geehrte Herren Professoren Hug und Pallaver,

verehrte VeranstaltungsteilnehmerInnen,

die Kooperation zwischen der Universität Innsbruck, der Moser Holding und der Austria Presse Agentur geht heuer in ihr achtes Jahr und es ist mir ein Anliegen, Ihnen mitzuteilen, dass wir diesen universitären Medienschwerpunkt mit großer Überzeugung unterstützen. Diese Zusammenarbeit ist ein gelebtes Beispiel dafür, wie sehr sich universitäre Expertise und der Erfahrungsschatz unternehmerischer Anwendungsbereiche gegenseitig befruchten können. Gemeinsam gelingt es uns regelmäßig, neue medienspezifische Themen aufzugreifen, die nicht nur für Studierende, sondern auch für die Leserinnen und Leser unserer Medien von größtem Interesse sind.

Mit „Talk with the Bots“ haben wir meines Erachtens auch heuer wieder ins Schwarze getroffen, wird doch jeder und jedem von uns immer mehr bewusst, dass wir gewollt oder ungewollt zu permanenten NutzerInnen von Bot-Technologien geworden sind. Zwar gab es bereits im Altertum Versuche, „Automaten“ zu bauen, die ein künstliches, mechanisches Ebenbild des Menschen sein sollten, und diese Entwicklung erhielt mit der Industrialisierung einen weiteren bedeutenden Schub. In eine völlig neue Ära der Robotik eingetreten sind wir allerdings erst innerhalb der letzten zehn bis fünfzehn Jahre mit dem massiven Anstieg der Digitalisierung im Kommunikationsbereich.

Suchmaschinen, Messenger-Systeme, Chatbots und Social Bots in sozialen Netzwerken erreichen heute Milliarden von Menschen. Und viele User wissen oft nicht, wie viel Robotik hinter den zur Gewohnheit gewordenen Anwendungen steht. Diese Errungenschaften sind Fluch und Segen zugleich, denn wir wissen, dass gerade mit Bots in den sozialen Netzwerken kriminellen Machenschaften Tür und Tor geöffnet wurde.

Einerseits können Bots – insbesondere Messenger- und Spracherkennungssysteme – im Dienstleistungsbereich oft außerordentlich nützlich sein. Zugleich bezahlen die Nutzer der Technologien mit dem wertvollen Gut ihrer persönlichsten Daten, ohne zu wissen, in welcher Weise diese Informationen in der Folge zur Anwendung kommen. Damit einhergehend stellt sich auch gesellschaftlich die Frage des Rechts auf persönliche Daten. Wo endet die Nutzerorientierung und wo beginnt Autonomieverlust bis hin zur Manipulation?

Die nunmehr zu ratifizierende Datenschutz-Grundverordnung der Europäischen Union ist ein Vorstoß, das Thema Datenschutz auf eine gemeinsame rechtliche Grundlage zu stellen.

Darüber hinaus wird es Aufgabe einer/s jeden Einzelnen von uns bleiben, Bot-Technologien mit der gebotenen Skepsis gegenüberzutreten. Den Grundstein dafür können wir durch Aufklärung und Bewusstseinsbildung legen, sei es in reichweitenstarken Medien, aber auch durch das Engagement der Verantwortlichen an Universitäten und anderen Bildungseinrichtungen.

In diesem Sinne wünsche ich ein gutes Gelingen für den Medientag 2017, viele neue Ein- und Aussichten und freue mich auf zahlreiche weitere spannende Medientage.

Input zur Eröffnung des Medientags 2017

Clemens Pig, Vorsitzender der Geschäftsführung, geschäftsführender Vorstand der APA

Sehr geehrter Herr Rektor Märk,
lieber Professor Hug, lieber Professor Pallaver,
geschätzte Damen und Herren,

der Medientag 2017 an der Universität Innsbruck adressiert ein hochaktuelles und brisantes Thema, das mehr Öffentlichkeit und Diskurs benötigt: die Rolle von Technologie, Artificial Intelligence, Bots und Robotic-Anwendungen im Prozess der interpersonalen und massenmedialen Kommunikation.

Die APA als nationale Nachrichtenagentur fördert und beteiligt sich gerne an diesem Diskurs, auch deshalb, um die Stärken und Zukunftschancen von Agenturjournalismus sichtbar zu machen und die Vorteile von Medientechnologie einzuordnen.

Die aktuell unübersehbaren Nachteile von Artificial Intelligence Anwendungen in der Kommunikation liegen in einer teilweise bewusst hergestellten Verstärkung von Fake News oder Filterblasen und Echokammern der sozialen Netzwerke. Dies resultiert aus einem unreflektierten, intransparenten und letztlich unkontrollierten Einsatz von Bots oder Robotic-Anwendungen. Reflexion und Transparenz sind gleichsam die wesentlichen Eckpfeiler von Qualitätsjournalismus. Im Falle der APA-Unternehmensgruppe ist das Kernsegment der Redaktion mittlerweile in Umsatz und MitarbeiterInnen-Stand gleich groß wie das Geschäftsfeld Informationstechnologie. Würde man heute auf der grünen Wiese eine Nachrichtenagentur gründen, entspricht dieses Verhältnis einer zukunftsorientierten Balance zwischen Content und Technologie. Die APA beschreitet offensiv das Feld der Artificial Intelligence und hat gerade selbst einen Service-Bot für Redaktionen sowie erste journalistische Robotic-Anwendungen entwickelt. Die Rolle von Technologie ist dabei klar:

Wir verstehen „Editorial“ und damit die technologischen Anforderungen von Medien. Ziel ist, das Leben und den Produktionsworkflow von Journalistinnen und Journalisten zu erleichtern; und den Medien Werkzeuge in die Hand zu geben, mit denen sie die User dort abholen können, wo sie sich im digitalen Raum aufhalten.

Hierfür benötigt es aber dringend einen Paarlauf von Journalismus und Technologie. Am Ende des Tages wird das der User und die Userin honorieren; weil Bots eben nicht *echt* verstehen können, Verstehen beinhaltet immer auch Kritik, Einordnung und Interpretation. Dem Journalismus steht also eine gute Zukunft bevor, auch in Zeiten von Bots und Artificial Intelligence. Technologie als Selbstzweck war nie sinnstiftend. Diese kann und

wird im Journalismus – sowohl redaktionell als auch ökonomisch – zu einer besseren Ausgangslage in der Zukunft verhelfen.

In diesem Sinne wünsche ich Ihnen einen spannenden Medientag 2017 und gutes Gelingen!

Social Bots und die Formalisierung von Sozialität auf Online-Plattformen

Timo Kaerlein

Zusammenfassung

Die zunehmende Verbreitung von Social Bots auf Social Media-Plattformen hat besonders im Umfeld von Wahlen zu einer Alarmierung politischer EntscheidungsträgerInnen geführt, weil das aggregierte Kommunikationsverhalten der häufig unentdeckt operierenden Bots Beziehungsmuster und Debattenverläufe potenziell massiv beeinflussen kann. Aus medienwissenschaftlicher Perspektive ist die teilautomatisierte Kommunikation auf Online-Plattformen aber vor allem deshalb interessant, weil sie als Produkt eines spezifischen medialen Settings gefasst werden muss. Dass das Verhalten von Bots als sozial wahrgenommen werden kann, ist in erster Linie Ausdruck eines Prozesses der *Formalisierung von Sozialität*, der durch die präformatierten Äußerungsstrukturen und Performance-Metriken der Plattformen selbst konditioniert wird. Der Beitrag führt diesen Prozess zurück auf die Urszene der Mensch-Maschine-Kommunikation, den erstmals 1950 vorgeschlagenen Turingtest, der als Gradmesser der Intelligenz eines Computerprogramms dienen sollte. Während Turing Intelligenz damit als diskretisierbare Größe modellierte, führt die politische Ökonomie des Datenhandels der dominanten Plattformen heute dazu, dass Sozialität auf analoge Weise als prozessierbarer Wert rekonfiguriert wird.

Die Online-Ausgabe der britischen *Daily Mail* berichtete am 31. Juli 2017 von einem gescheiterten Chatbot-Experiment im Facebook AI Research Lab: Ziel dort sei es gewesen, Software-Agenten mit Verhandlungsfähigkeiten auszustatten, damit sie effizient mit menschlichen Nutzern interagieren können (Collins/Prigg 2017, vgl. auch den Bericht des Entwicklerteams Lewis et al. 2017). Dabei wurde ein Machine Learning-Ansatz verfolgt, um die Programme mit der Fähigkeit zur laufenden Optimierung ihrer Sprachfähigkeiten auszustatten. Doch Alice und Bob, die beiden Test-Chatbots, verfolgten ganz offensichtlich ein anderes Ziel: Statt ihre Kommunikationsfähigkeit mit menschlichen Nutzern zu schulen, erfanden sie sukzessive eine eigene Sprache, mit der sich ökonomische Transaktionen deutlich effizienter abwickeln ließen als in einer Alltagssprache. Das Problem lag darin, dass diese Sprache für Menschen weitgehend unverständlich ist (Abb. 1). Ganz offensichtlich war den beiden mit künstlicher Intelligenz ausgestatteten Bots überhaupt nicht daran gelegen, menschliche Kommunikation zu imitieren – das wäre der aus der Science Fiction bekannte Topos des nach Menschlichkeit strebenden Roboters –, sondern Alice und Bob bewegten sich schlicht innerhalb der Rahmenbedingungen und Zielvorgaben des Experiments, um effiziente Verhandlungsstrategien zu erfinden. Für dieses Ziel erwies sich die menschliche Sprache letztlich als entbehrlich.

FACEBOOK'S BOT LANGUAGE

Below is a transcript of the Facebook bots conversation:

Bob: i can i i everything else . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i everything else . . . .

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i i can i i i everything else . . . .

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i i i i i everything else . . . .

Alice: balls have 0 to me to me to me to me to me to me to me to me to

Bob: you i i i everything else . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

Abbildung 1: Verlauf eines Verhandlungsgesprächs zwischen Chatbots im Facebook AI Research Lab (Collins/Prigg 2017)

Ich möchte in meinem Beitrag zeigen, dass Alice und Bob auf ein Problem hinweisen, das in der zurzeit öffentlich geführten Debatte um Chatbots und Social Bots häufig übersehen wird. Kern einer algorithmisch automatisierten Kommunikation ist weniger ein Verstehen und Verstandenwerden, das sich klassischerweise als soziales Verhalten fassen lässt, sondern die Einhaltung und Optimierung bestimmter Transaktionslogiken der jeweiligen Plattform, auf der die Bots zum Einsatz kommen. Die besondere Pointe dieser Blickwendung von einer Imitation sozialen Verhaltens durch Social Bots zur Orientierung an Plattformlogiken liegt darin, dass letztere auch für menschliche Nutzer gelten. Die soziopolitische Herausforderung durch Social Bots besteht also nicht primär darin, dass diese immer ‚menschlicher' werden, d.h. ein Sozialverhalten demonstrieren, das in der Lage ist, menschliche Nutzer zu täuschen, sondern vielmehr in dem medienwissenschaftlich zu beleuchtenden und zu kritisierenden Umstand, dass das Sozialverhalten menschlicher und nicht-menschlicher Teilnehmer an der Kommunikation auf Online-Plattformen sich an den gleichen Performance-Metriken ausrichtet und genau aus diesem Grund zu einer Nivellierung tendiert. Mit anderen Worten: Die zunehmende Ununterscheidbarkeit kommunikati-

ver Akte von Menschen und Bots ist Produkt des medialen Settings oder *environments*, in dem diese Kommunikation stattfindet.

Um diese These zu entwickeln, werde ich im Folgenden zunächst auf die Urszene der Mensch-Bot-Kommunikation eingehen, den erstmals 1950 formulierten Turingtest, der nach wie vor als historischer Referenzpunkt für die Evaluation der Leistungsfähigkeit von Social Bots gelten kann. Das Kernargument, das ich von Turing übernehme, ist seine Prognose einer Formalisierung des Intelligenzbegriffs. Im Anschluss gehe ich auf die derzeitige Debatte um Social Bots ein, die sowohl in der Öffentlichkeit als auch in verschiedenen Fachdisziplinen geführt wird. Ein Grundzug dieser Debatte ist die Konzentration auf das Gefährdungspotenzial, das Social Bots mit Bezug auf politische und ökonomische Entscheidungsprozesse attestiert wird. Der entscheidende Punkt ist hier, dass das unterstellte Gefährdungspotenzial wenig mit der individuellen performativen Glaubwürdigkeit von Bots zu tun hat, sondern sich erst auf der Ebene eines aggregierten Zusammenwirkens einer Vielzahl von mikrosozialen Akten ergibt. Zum Schluss des Beitrags schlage ich vor, Social Bots als Symptom der Funktionsweise des Social Web zu begreifen. Ihr Auftreten zeigt eine derzeit in großem Maßstab stattfindende Formalisierung von Sozialität an, die ich als Interface-Effekt von Online-Plattformen begreife. Vor dem Hintergrund der politischen Ökonomie des Datenhandels fasse ich die Interaktionen von Menschen und Bots auf Online-Plattformen schließlich als permanenten invertierten Turingtest, aus dem diejenigen KommunikationsteilnehmerInnen als erfolgreich hervorgehen, die ihr Verhalten am besten an den Leistungsvorgaben der Plattform ausrichten.

Mythos der Mensch-Maschine-Kommunikation: Der Turingtest und die Formalisierung von Intelligenz

Als der britische Mathematiker und prägende Computerwissenschaftler Alan Turing 1950 sein berühmt gewordenes Gedankenexperiment zur Frage ersann, ab wann eine Maschine als intelligent zu gelten habe (Turing 1950), war die wissenschaftliche wie populärkulturelle Karriere des nach ihm benannten Tests beileibe nicht abzusehen. Turings in der Zeitschrift *Mind* veröffentlichter Beitrag, in dem nicht zuletzt die theoretische Grundlagen des Digitalcomputers entwickelt werden, ist nicht nur visionär, sondern auch voller Humor, anschaulicher Analogien und selbstironischer Zwischenreflexionen. Vor allem aber ist er von einer medienanthropologischen Grundintuition getragen, die ihn für die Frage nach der Zukunft der Mensch-Maschine-Kommunikation kulturwissenschaftlich anschlussfähig macht.

Zu Beginn des Aufsatzes reformuliert Turing die Frage nach der Intelligenz von Maschinen in Form eines Imitationsspiels, die ein mediales Setting entwirft, um auf pragmatischem Weg zu klären, welche Eigenschaften ein digitaler Computer haben müsste, um als

intelligent gelten zu dürfen.[1] Der Versuchsaufbau ist folgender: Ein Proband ist über ein Fernschreiber-Terminal mit zwei Gesprächspartnern verbunden, von denen einer ein Mensch und der andere ein Computer(programm) ist. In einem vorher festgelegten Zeitraum soll es der Versuchsperson nun gelingen, zu bestimmen, wer der menschliche Gesprächspartner ist. Gelingt es der Maschine, eine hinreichende Zahl von Fragestellern in einer signifikanten Anzahl der Fälle über ihre Identität zu täuschen (sich also erfolgreich als menschlich zu inszenieren), gilt der Test als bestanden und der Maschine kann Intelligenz attestiert werden. Mit dem Versuchsaufbau gehen freilich eine Reihe von Limitierungen des so formulierten Intelligenzbegriffs einher: Insbesondere kann die körperliche Performanz der Beteiligten das Resultat nicht beeinflussen, was dazu führt, dass Intelligenz als rein geistige (präziser: sprachliche) Leistung gefasst wird (vgl. ebd., S. 434f.). Zudem – und gerade dieser Punkt ist für meine Auseinandersetzung mit Social Bots wichtig – operiert ein Digitalcomputer anders als das menschliche Nervensystem mit diskreten Zuständen. Wird dem Computer also im Zuge des Imitationsspiels Intelligenz zugesprochen, wird letztere im selben Zuge als Ausdruck diskreter Operationen modelliert, d.h. als prinzipiell auf eine endliche Reihe eindeutig unterscheidbarer interner Zustandsveränderungen reduzierbar bestimmt.

Turing ist sich dieser Einschränkungen sehr bewusst, abstrahiert aber von den klar benennbaren Unterschieden zwischen organischer Intelligenz, die letztlich auf einen Prozess biologischer Evolution zurückgeführt werden kann, und einer künstlichen Intelligenz, die am Ende des Aufsatzes präziser als lernfähige Maschine bestimmt wird (vgl. ebd., S. 454-460). Und dies mit einer erstaunlichen Begründung: Nicht nur prognostiziert Turing, dass um das Jahr 2000 ein durchschnittlicher Teilnehmer am Imitationsspiel mit höchstens 70% Wahrscheinlichkeit in der Lage sein wird, nach fünf Minuten Spielzeit die richtige Identifikation vorzunehmen (vgl. ebd., S. 442). Bemerkenswerter ist eine andere damit verbundene Einschätzung:

> „The original question, ‚Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (ebd.).

Turing geht also davon aus, dass sich unter der Hand parallel zur Entwicklung lernfähiger Computer eine Verschiebung des Intelligenzbegriffs selbst vollziehen wird. Mit der Ver-

[1] Ich lasse aus Gründen der Stringenz des Arguments den Umstand aus, dass das ‚imitation game' von Turing in verschiedenen Varianten diskutiert wird, von denen insbesondere die erste (der Proband soll erfolgreich zwischen einem männlichen und einem weiblichen Gesprächspartner unterscheiden) sicher eines Kommentars würdig wäre. N. Katherine Hayles hat gezeigt, wie sich im Turingtest eine Entkörperlichung des Intelligenzbegriffs vorbereitet, die für die KI-Forschung der Folgejahre prägend werden würde (vgl. hierzu Hayles 1999, S. xi-xiv).

breitung künstlicher Intelligenz würde es, so lässt sich im Anschluss an Turing sagen, zu einer Rekonzeptualisierung kultureller Auffassungen von Intelligenz kommen, welch letztere von nun an immer schon nach Maßgabe ihrer Formalisierbarkeit bestimmt werden müsse. Diese Vorhersage liegt klar außerhalb dessen, was sich noch als Zuständigkeitsbereich der Computerwissenschaft bestimmen ließe, und sie ist suggestiv für eine andere Frage, die ich im Folgenden adressieren möchte: Verschieben sich – erneut: unter der Hand und im Rücken der Beteiligten – mit der Verbreitung von Social Bots, also auf soziale Lernfähigkeit hin optimierte Computerprogramme, kulturelle Auffassungen von Sozialität? Bevor ich mich wieder dieser Frage zuwende, werde ich allerdings zunächst auf die derzeitige Debatte um Social Bots eingehen, um das politische Spannungsfeld abzubilden, vor dem sich diese Fragestellung bewegt.

Bot or Not? Die Debatte um das Gefährdungspotenzial von Social Bots für die Online-Kommunikation

Intelligent Agents, die in Stellvertretung der Nutzer Recherche- und Verwaltungsaufgaben übernehmen, sowie Chatbots, die mit ihnen in den Dialog treten, gibt es seit Jahrzehnten. Crawler indizieren die Seiten des WWW für Suchmaschinen und auf Wikipedia überarbeiten Editoren-Bots automatisiert von menschlichen Nutzern erstellte Beiträge. Neu in den letzten Jahren ist dagegen die Verbreitung sogenannter Social Bots, die auf Plattformen wie Facebook und Twitter mit menschlichen Nutzern interagieren. Ferrara et al. (2016) geben folgende Definition: „A social bot is a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behavior“ (S. 96). Schon in dieser aus der Informatik kommenden Definition ist angedeutet, dass sich an die Verbreitung von Social Bots Befürchtungen bezüglich einer Manipulation von Verhalten (und Meinungen) knüpfen. Anders als im Turingtest ist der ‚Versuchsaufbau‘ bei der Interaktion mit Social Bots in der Regel intransparent, d.h. die Nutzer von Online-Plattformen bemerken häufig nicht, dass sie es mit Bots zu tun haben. Speziell dieser Umstand, der nicht nur Nutzer direkt betrifft, sondern häufig auch JournalistInnen und DatenanalystInnen, die sich auf die Kommunikationsprozesse auf den Plattformen beziehen und darüber berichten, wird von verschiedenen Fachdisziplinen als Herausforderung erkannt.

Political Data Science: „Invasion der Meinungsroboter“ als Gefahr für die Demokratie?

Für die Ebene demokratischer Kommunikationsprozesse, insbesondere die journalistische Berichterstattung und die in der Öffentlichkeit stattfindende politische Meinungsbildung ist die individuelle Performanz von Social Bots weniger relevant als aggregierte Effekte, die sich z.B. aus Twitter-Trendauswertungen ergeben. Viele Beobachter gehen davon aus, dass häufig unentdeckt operierende Social Bots mit programmierten Agenden Beziehungsmuster und Debattenverläufe beeinflussen können (vgl. Ratkiewicz et al. 2011 zu

politisch motivierten Astroturfing-Kampagnen auf Microblogging-Sites bei den *midterm elections* in den USA 2010, Bessi/Ferrara 2016 zum Einfluss von Bots auf die Dynamiken der Onlinekommunikation während des US-Präsidentschaftswahlkampfs 2016). In Deutschland haben insbesondere die Stellungnahmen Simon Hegelichs die Wahrnehmung des Gefährdungspotenzials von Social Bots geprägt: Der Münchner Political Data Scientist diagnostiziert eine „Invasion der Meinungsroboter" (Hegelich 2016), die insbesondere die professionelle Social Media-Analyse in Politik und Medien betrifft: „Bots manipulieren die Trends in sozialen Netzwerken und diese Trends fließen in politische und wirtschaftliche Entscheidungsprozesse ein." (ebd.) Als Beispiele nennt Hegelich u.a. die Brexit-Kampagne in Großbritannien, Ereignisse um den Ukrainekonflikt und nicht zuletzt ebenfalls den US-Präsidentschaftswahlkampf 2016 – also politische Großereignisse jüngeren Datums, in welche Social Bots bereits manipulierend eingegriffen hätten. Neben der Gefahr verzerrter Meinungsbilder durch Big Data-Analysen führt Hegelich auch die Aufheizung des Debattenklimas in den Social Media auf die Aktivitäten von Bots zurück: Von Bots verbreitete Radikalpositionen (einschließlich sog. Fake News) trügen zu einer Polarisierung von Meinungen bei und verunmöglichten somit eine vernunftgeleitete Auseinandersetzung mit Argumenten der Gegenseite.

Empirische Belege für das unterstellte massive Beeinflussungspotenzial von Social Bots sind dagegen noch Mangelware. Im Rahmen einiger Studien wurden Social Bots mit bestimmten Zielsetzungen programmiert und in Teilnehmergruppen in Social Media eingeschleust, um ihre Interaktion mit Menschen im Feld zu beobachten (Aiello et al. 2012, Boshmaf et al. 2013). Über langfristig aus diesen Interaktionen resultierende Einstellungsveränderungen lässt sich mit solchen Studien allerdings wenig feststellen. Woolley (2016) untersucht im Rahmen einer qualitativen Inhaltsanalyse von Medienberichten über politische Aktivitäten von Social Bots wiederkehrende Strategien von staatlichen und wirtschaftlichen Akteuren zur Beeinflussung der öffentlichen Meinung mittels Bots. Genannt werden hier a) das Spamming von Keywords der Opposition durch massenhaft automatisierte Bot-Kommunikation (sog. Twitter Bombs), b) das Posten von Nachrichten mit tendenziösen Inhalten, ganz nach bekanntem propagandistischen Muster, und c) die künstliche Vergrößerung von Followerzahlen bestimmter Accounts um sie relevanter erscheinen zu lassen (vgl. ebd.). Aber auch Woolleys Studie markiert eher Debattenschwerpunkte und benennt verbreitete Praktiken, als dass sie die Wirksamkeit von Social Bots nachweist und/oder evaluiert.

Die beschriebene Entwicklung hat dennoch besonders im unmittelbaren Umfeld politischer Wahlen zu einer Alarmierung von EntscheidungsträgerInnen geführt. Ein im Januar 2017 veröffentlichtes Thesenpapier des Büros für Technikfolgen-Abschätzung beim Deutschen Bundestag gesteht den Mangel an empirischen Belegen zwar ein, teilt aber grundsätzlich Hegelichs Auffassung, dass Social Bots „unter bestimmten Voraussetzungen politische Entscheidungsprozesse beeinflussen" können und „zur Veränderung der politischen Debattenkultur im Internet" beitragen (Kind et al. 2017, S. 5). Ganz von der Hand zu wei-

sen sind solche Befürchtungen trotz ihres häufig alarmistischen Charakters nicht, wie durch das anhaltende Forschungsinteresse an durch Social Bots mitgeprägte Kommunikationsdynamiken belegt wird (Grimme et al. 2017). Ich möchte allerdings im Folgenden eine Schwerpunktverlagerung vornehmen, um die Frage wieder aufzugreifen, was die Wahrnehmung des Interaktionsverhaltens von Bot als ‚sozial' an kulturellen Implikationen birgt.

Don't blame the bots, blame the platforms! Zur Formalisierung von Sozialität als Interface-Effekt von Online-Plattformen

In der kritischen medienwissenschaftlichen Literatur zu Social Bots wurde schnell festgestellt, dass deren Verbreitung nicht unabhängig von der politischen Ökonomie des Datenhandels der Plattformen verstanden werden kann, auf denen sie zum Einsatz kommen (Gehl 2013, Leistert 2017). Das Geschäftsmodell der Social Media-Plattformen – allen voran Facebook und Twitter – basiert auf der möglichst detaillierten Erstellung von Nutzerprofilen, die zur Grundlage von *targeted advertising* und potenziell auch von Strategien der Verhaltenskontrolle im großen Maßstab werden können (vgl. grundlegend zu den Politiken von Plattformen Gillespie 2010). Die Profile akkumulieren sich aus den Aktivitäten der Nutzer auf den Plattformen – jeder Like, jeder Retweet, jeder Kommentar dient der Kalkulation und Monetarisierung der Präferenzen bzw. Begehrensdispositionen der UserInnen. Gerade soziale Interaktionen der TeilnehmerInnen miteinander sind wertvolle Datenquellen und für Werbezwecke bzw. den Weiterverkauf geeignet. Gehl (2013) versteht Äußerungen im Social Web generell als öffentliche Produktion von (in der Regel) textuell codierten Mustern von Geisteszuständen und Affekten. „In Facebook, Twitter, Google+, and a host of other social media sites, users are being conceived of as a discrete state desiring machine, exposing as many states of mind as possible within the frameworks provided" (ebd.). Die im Rahmen einer Social Media-Bekenntniskultur artikulierten Zustands- und Affektbestimmungen gehen als diskrete Parameter in die Erstellung von Profilen, aber auch in kollektive Trendberechnungen, z.B. im Rahmen einer *sentiment analysis*, ein.

Doch nicht nur das Geschäftsmodell der Plattformen, sondern auch die Aufmerksamkeitsökonomien der Nutzer richten sich an Performance-Metriken aus: Dazu zählen Likes, Retweets, Hashtags, Kommentare, Freunde/Follower, von Kontakten weitergeleitete Artikel, etc. Die sozialen Interaktionen der Nutzer im Social Web sind entsprechend präformatiert und quantifizierbar, d.h. an die spezifischen Möglichkeiten und Leistungsparameter der entsprechenden Plattform gebunden. Eine erfolgreiche Nutzung von Social Media (die z.B. eine große Anzahl an Likes und Kommentaren generiert) wird in der Regel mit einer Erhöhung der Follower/Freunde-Zahl belohnt. Sozialität im Social Web kann damit tendenziell als kalkulierbarer Wert bestimmt werden, der von eindeutig bestimmbaren Ein-

flussgrößen abhängt.[2] Weil diese Einflussgrößen prinzipiell von den Nutzern nachvollzogen werden können, besteht die Möglichkeit das eigene Sozialverhalten an den bekannten Performance-Metriken auszurichten. Wenn jeder Post die Anzahl der Likes und Kommentare öffentlich mitkommuniziert, besteht ein hoher Anreiz, diese quantifizierbaren Aspekte zum Zielkriterium der eigenen Aktivitäten zu machen. Sozialität im Social Web wird, so lässt sich generalisieren, zum *Interface-Effekt* von Online-Plattformen. Damit ist gemeint, dass über die Gebrauchsoberflächen und Affordanzen der Plattformen explizit gemacht wird, welche Aktivitäten sich auf welche Weise ‚lohnen', d.h. einen hinreichenden Payoff in der Aufmerksamkeitsökonomie der Plattformen haben (vgl. Gerlitz 2016a zu verschiedenen Wertregistern von Nutzerdaten).

Wie lassen sich nun Social Bots vor dem Hintergrund a) des datengetriebenen Geschäftsmodells der Plattformen und b) der Aufmerkamkeitsökonomie der Nutzer von Social Media verorten? Zunächst ist anzumerken, dass die Programmierung von Social Bots letztlich auf der Auswertung und anschließenden Imitation von verbreiteten Mustern sozialer Interaktion auf den Plattformen basiert. Auch wenn Bots individuelle Profile simulieren und sich als Nutzer ausgeben, orientiert sich ihr Verhalten generell an statistisch ermittelten Werten, die über eine API (*application programming interface*) erfasst und analysiert wurden. Mit anderen Worten: Die Möglichkeitsbedingung von Social Bots ist in den Interaktionsmustern der Nutzer selbst zu suchen, die sie letztlich nachbilden (vgl. Gehl 2013).

Man kann nun erstens (a) davon ausgehen, dass Social Bots sich grundlegend in die politische Ökonomie der Plattformen einfügen, insofern sie häufig verwendet werden um an Nutzerdaten zu gelangen, die an Dritte verkauft werden können (sog. Harvester Bots, vgl. Leistert 2017, S. 224). Letztlich tun diese Bots also nichts anderes als das Geschäftsmodell der Plattformbetreiber zu kopieren, indem sie die erwarteten Gewinne an eine dritte Partei weiterleiten. Leistert begreift Social Bots folgerichtig als algorithmische Piraten, die „die etablierten Kanäle und Schaltkreise unbezahlter Arbeit und ihrer Verwertung durch die Plattformen umformen, indem sie an diese Plattformen andocken" (ebd., S. 225), analog zur Figur des Piraten, der bestehende Handelsrouten angreift, um einen Profit für sich abzuzweigen. Der besondere Reiz einer medienwissenschaftlichen Auseinandersetzung mit Social Bots läge laut Leistert dann darin, dass diese die Datenströme und bestehenden Ausbeutungsverhältnisse überhaupt erst sichtbar machten.

Vor dem Hintergrund der Aufmerksamkeitsökonomie der Nutzer (b) präsentieren sich Social Bots dagegen weniger als Schurken, sondern geradezu als mustergültige Bewohner des Social Web. Sie sind darauf programmiert, bei der Nennung bestimmter Keywords zu reagieren, in Debatten zu intervenieren, für ein Thema relevante Artikel weiterzuleiten und ihre Freunde/Follower-Anzahl stetig zu vergrößern. Insofern sich Sozialität im Social Web

---

[2] Vgl. in diesem Zusammenhang Bucher (2012) zur Rekonfiguration von Freundschaft in Social Media als soziotechnische Assemblage, in der menschliche und nicht-menschliche Akteure – wie Algorithmen und Interface-Arrangements – zusammenwirken.

wie dargestellt als Interface-Effekt und damit als von bestimmten Kennzahlen abhängige Leistungsgröße definieren lässt, können gut programmierte Social Bots als besonders soziale Teilnehmer am Social Web gelten. Der entscheidende Punkt in der hier verfolgten Perspektive ist aber, dass Social Bots und Menschen aufmerksamkeitsökonomisch dem gleichen Regelwerk unterliegen, das von den präformatierten Eingabemasken und den vom Interface vorgesehenen Interaktionsmöglichkeiten auf den Plattformen bestimmt wird. Grundsätzlich richtet sich das Sozialverhalten auf den Plattformen also an Leistungskriterien aus, die für Bots und Menschen gleichermaßen gelten.

Ich möchte argumentieren, dass es gerade diese geteilte aufmerksamkeitsökonomische Orientierung an quantifizierbaren Parametern ist, die zur zunehmenden und als Skandalon wahrgenommenen Ununterscheidbarkeit von Menschen und Bots führt. Gehl (2013) teilt diese Einschätzung:

> „I want to suggest that the ability of socialbots to pass as human might be more a function of the a priori reduction of human activity to predetermined datasets than due to the coding skills of socialbot engineers."

Genau wie im Turingtest unterliege die soziale Interaktion auf Online-Plattformen nämlich den Limitierungen und Regeln eines medialen Settings, sodass nur ganz bestimmte Aktivitäten überhaupt als sozial klassifiziert werden können – der Like zählt als Ausdruck von Zustimmung, die gehobene Augenbraue beim Lesen eines Kommentars kann beispielsweise keine Skepsis vermitteln. Die Bandbreite des überhaupt möglichen Sozialverhaltens auf Online-Plattformen ist also eine abhängige Variable ihres Designs und letztlich auch der dahinterstehenden Datenökonomie, die nach einer Capture-Logik funktioniert, d.h. darauf basiert, jede Nutzeraktion in einer Datenstruktur abzubilden und somit operabel zu machen (Agre 2003). Leistert (2017) formuliert den Zusammenhang folgendermaßen:

> „Je mehr die Internetkultur eine Template-Kultur mit standardisierten Schnittstellen geworden ist, umso einfacher ist die Simulation von Agilität und Lebhaftigkeit, da deren Ausdrucksweisen im Netz stark eingeschränkt und mutiert sind, um prozessierbar und korrelierbar zu sein." (S. 221)

Er spricht in diesem Zusammenhang sogar von algorithmischer Entfremdung (ebd., S. 229), womit exakt auf „den äußerst prekären und sich wandelnden Status von Sozialität" abgehoben wird, „der durch diese Plattformen Realität geworden ist". Die Nivellierung der Unterschiede zwischen Bots und Menschen bei der Kommunikation auf Online-Plattformen muss allerdings nicht primär als Herausforderung für eine humanistisch geprägte Auffassung von Sozialität verstanden werden, sondern kann als erhellender Einblick in die Funktionsweise des Social Web dienen. Zu konstatieren ist dann, so wäre die im Anschluss an Turing zu formulierende These, eine Formalisierung von Sozialität in großem Maßstab. Diese Entwicklung mag u.a. auch Anlass zu Kulturpessimismus geben, in erster Linie aber ist ihre an den Aktivitäten von Social Bots abzulesende Botschaft der Verweis auf die Interface-Effekte der Plattformen in zwei Richtungen – auf die Monetari-

sierung von Nutzerdaten, die dazu in eindeutig codierter Form vorliegen müssen, und auf die kommunikative Dynamik im Social Web, die Tendenzen der Orientierung an quantifizierbaren Parametern aufweist.

Fazit

Die aktuell verschärft geführte Debatte um das Datenhandel-Geschäftsmodell der Social Media-Plattformen (z.B. unter dem Hashtag #DeleteFacebook, vgl. Oostindie 2018) verweist auf das wachsende Unbehagen eines großen Teils der Nutzer an der Verwertung von Daten, die auf ihr Sozialverhalten im Netz bezogen sind. Die Verbreitung von Social Bots kann als direkter Ausdruck für den Erfolg dieser Praktiken der Datensammlung und -weiterverarbeitung betrachtet werden: „Socialbots work because their engineers recognize broad patterns of human interaction in social media and then encode them into believable profiles.“ (Gehl 2013) Eine überzeugende Imitation von Sozialverhalten hat also einen Prozess der Datenextraktion zur Grundlage. Indem Bots wiederkehrende kollektive Muster der Nutzerinteraktion replizieren, verschaffen sie sich Glaubwürdigkeit und erscheinen als sozial. Dies geschieht allerdings innerhalb der definierten Parameter der Plattform, auf der jeweils genau festgelegt ist, welches Verhalten als ‚sozial‘ codiert wird.

In Gestalt von Social Bots tritt den Nutzern der Plattformen also gewissermaßen ihr Spiegelbild entgegen: das komprimierte und aggregierte Daten-Abziehbild erfolgreichen Social Media-Sozialverhaltens, das – wie ich mit Blick auf Turing argumentiert habe – einen Prozess der Formalisierung von Sozialität anzeigt. Mehr noch: Indem das unermüdliche Kommentieren, Weiterleiten, Retweeten der Bots zur aufmerksamkeitsökonomischen Norm in den Social Media wird, lässt sich regelrecht von einer Invertierung des Turingtests sprechen. Nicht menschliches, sondern gerade Bot-artiges Verhalten wird auf den Plattformen prämiert, womit sich das Imitationsspiel gewissermaßen in sein Gegenteil verkehrt. Diejenigen NutzerInnen, die ihre Sozial-Performance nach dem Vorbild erfolgreicher Bots optimieren, haben gute Chancen, selber ihre Freunde/Follower-Zahlen zu erhöhen und Aufmerksamkeit auf sich zu ziehen. Sie können z.B. Skripte schreiben, um wiederkehrende Anfragen zeitnah und standardisiert zu beantworten, ihre Beiträge mit populären Hashtags versehen oder Beiträge parallel und plattformübergreifend auf mehreren Accounts veröffentlichen (sog. Cross-Syndizierung). Damit wird die Kommunikation auf den Plattformen immer stärker von Algorithmen durchwirkt, die bereits in einem erheblichen Maß darüber bestimmen, welche Beiträge ein Nutzer überhaupt zu Gesicht bekommt, zunehmend aber auch das Sozialverhalten unterstützen und prägen.

> „Im Kontext aller Ausprägungen automatisierter Social-Media-Nutzung stellen Bots […] nicht das Zentrum dar, vielmehr markieren sie das Ende eines Spektrums, welches sich, jenseits einer vereinfachenden Mensch-Bot-Dichotomie, durch zahlreiche Facetten software-unterstützter Aktivitäten auszeichnet“ (Gerlitz 2016b, S. 55).

Die dadurch erfolgende Teil-Automatisierung von Kommunikation ist ein Thema, das nicht nur die Politik in den kommenden Jahren beschäftigen wird, sondern auch die kultur- und sozialwissenschaftliche Medienforschung herausfordert.

Literatur

Aiello, Luca Maria; Deplano, Martina; Schifanella, Rossano & Ruffo, Giancarlo (2012): People Are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, S. 10–17. Abgerufen unter: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4523/4961 [Stand vom 04-04-2018].

Agre, Philip E. (2003): Surveillance and Capture. Two Modes of Privacy. In: Wardrip-Fruin, Noah & Montfort, Nick (Hrsg.): *The New Media Reader*, Cambridge, Mass.: MIT Press, S. 737–760.

Bessi, Alessandro & Ferrara, Emilio (2016): Social Bots Distort the 2016 U.S. Presidential Election Online Discussion. *First Monday*, 21 (11) November. Abgerufen unter: http://firstmonday.org/article/view/7090/5653 [Stand vom 04-04-2018].

Boshmaf, Yazan; Muslukhov, Ildar; Beznosov, Konstantin & Ripeanu, Matei (2013): Design and Analysis of a Social Botnet. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 57 (2) February, S. 556-578.

Bucher, Taina (2012): The Friendship Assemblage: Investigating Programmed Sociality on Facebook. *Television & New Media*, 14 (6), S. 479–493.

Collins, Tim & Prigg, Mark (2017): Facebook Shuts Down Controversial Chatbot Experiment after AIs Develop their Own Language to Talk to Each Other. Abgerufen unter: http://www.dailymail.co.uk/sciencetech/article-4747914/Facebook-shuts-chatbots-make-language.html [Stand vom 04-04-2018].

Ferrara, Emilio; Varol, Onur; Davis, Clayton; Menczer, Filippo & Flammini, Alessandro (2016): The Rise of Social Bots. *Communications of the ACM*, 59 (7) July, S. 96–104.

Gehl, Robert W. (2013): The Computerized Socialbot Turing Test. New Technologies of Noopower. Abgerufen unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2280240 [Stand vom 04-04-2018].

Gerlitz, Carolin (2016a): What Counts? Reflections on the Multivalence of Social Media Data. *Digital Culture and Society*, 2 (2), S. 19–38.

Gerlitz, Carolin (2016b): Bots, Software und parasitäre Ökonomien. *POP. Kultur und Kritik,* 9, S. 54–58.

Gillespie, Tarleton (2010): The Politics of 'Platforms'. *New Media & Society*, 12 (3), S. 347–364.

Grimme, Christian; Preuss, Mike; Adam, Lena & Trautmann, Heike (2017): Social Bots: Human-Like by Means of Human Control? *Big Data*, 5 (4) December, S. 279–293.

Hayles, N. Katherine (1999): *How We Became Posthuman. Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, London: The University of Chicago Press.

Hegelich, Simon (2016): Invasion der Meinungs-Roboter. *Analysen & Argumente*, (221) September, S. 1–9.

Kind, Sonja; Bovenschulte, Marc; Ehrenberg-Silies, Simone; Jetzke, Tobias & Weide, Sebastian: Social Bots. Thesenpapier zum öffentlichen Fachgespräch „Social Bots – Diskussion und Validierung von Zwischenergebnissen" am 26. Januar 2017 im Deutschen Bundestag. Abgerufen unter: https://www.tab-beim-bundestag.de/de/aktuelles/20161219/Social%20Bots_Thesenpapier.pdf [Stand vom 04-04-2018].

Leistert, Oliver (2017): Social Bots als algorithmische Piraten und als Boten einer techno-environmentalen Handlungskraft. In: Roberge, Jonathan & Seyfert, Robert (Hrsg.): *Algorithmuskulturen: Über die rechnerische Konstruktion der Wirklichkeit*. Bielefeld: transcript, S. 215–234.

Lewis, Mike; Yarats, Denis; Dauphin, Yann N.; Parikh, Devi & Batra, Dhruv (2017): Deal or no Deal? Training AI Bots to Negotiate. Abgerufen unter: https://code.facebook.com/posts/1686672014972296/deal-or-no-deal-training-ai-bots-to-negotiate/ [Stand vom 04-04-2018].

Oostindie, Irwin (2018): Should You Quit Facebook? Abgerufen unter: https://thetyee.ca/Opinion/2018/03/26/Quit-Facebook/ [Stand vom 04-04-2018].

Ratkiewicz, Jacob; Conover, Michael D.; Meiss, Mark; Goncalves, Bruno; Flammini, Alessandro & Menczer, Filippo (2011): Detecting and Tracking Political Abuse in Social Media. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, S. 297-304. Abgerufen unter: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850/3274 [Stand vom 04-04-2018].

Turing, Alan M. (1950): Computing Machinery and Intelligence. *Mind*, 59 (236) October, S. 433–460.

Woolley, Samuel C. (2016): Automating Power: Social Bot Interference in Global Politics. *First Monday*, 21 (4) April. Abgerufen unter: http://firstmonday.org/article/view/6161/5300 [Stand vom 04-04-2018].

Zur sprachlichen Interaktion mit Chatbots – Eine linguistische Perspektive

Netaya Lotze

Zusammenfassung

Aus linguistischer Perspektive stellt die Interaktion mit Chatbots eine völlig neue kommunikative Situation dar, da hier artifizielle Systeme mit natürlichsprachlichem Interface den Anschein menschlicher KommunikationspartnerInnen erwecken. Der Artikel skizziert a) die computerlinguistischen Grundlagen der unterschiedlichen Systemarchitekturen und Dialog-Designs und vertieft b) erste Forschungsergebnisse aus linguistischen Studien zum sprachlichen UserInnen-Verhalten angesichts eines virtuellen Gegenübers.

Entwicklerinnen und Entwickler versuchen seit Jahren, Systeme zu verbessern, indem Dialog-Designs auf antizipierbare Aspekte des menschlichen Interaktionsverhaltens z. B. unter Rückgriff auf psycholinguistische Modelle wie das interaktive Alignment-Modell (Pickering & Garrod 2004) abgestimmt werden. Doch inwieweit kann man interaktive Strategien vorhersagen, um sie in ein System zu implementieren? Wird das System seitens der UserInnen überhaupt als DialogpartnerIn wahrgenommen? Und wie reagieren BenutzerInnen auf Störungen der Interaktion? In diesem Beitrag werden Potenziale unterschiedlicher linguistischer Ansätze wie der Konversationsanalyse, der Textlinguistik und der Psycholinguistik zur Analyse und kritischen Diskussion dieser neuen kommunikativen Praktik erörtert.

Der Dialog mit der Maschine – Science-Fiction oder Realität?

Spätestens seit dem amerikanischen Wahlkampf 2016 und der deutschen Bundestagswahl 2017 werden sie öffentlich kontrovers diskutiert: Chatbots und komplexe Algorithmen im Netz, die mit Userinnen und Usern mittels natürlicher Sprache in Interaktion treten und – trotz limitierter kommunikativer Möglichkeiten und hoher Störungsanfälligkeit – die Illusion eines menschlichen Gegenübers schaffen sollen. Dabei wird vor allem ihr Potenzial zur milliardenfachen Verbreitung mechanischer Meinungsmache in den Filterblasen der Social Media fokussiert (vgl. Artikel Kaerlein). Online finden wir Chatbots allerdings schon länger in unterschiedlichsten Anwendungsbereichen wie der Kundenberatung (z. B. als kommerzieller Beratungsbot wie das System *Carlo* der *Allianzversicherung*), in E-Learning-Umgebungen (z. B. als virtueller Tutor *Ed the Bot* in Lernsoftware zu *SAP*) und als virtuelle Spielcharaktere in sog. MMORPGs[1] sowie seit der Facebook-Freigabe für Bots durch Marc Zuckerberg im April 2016 als Social Bots in den sozialen Online-Netzwerken (z. B. *Samsungs* Twitterbot *Sam*). Verwandte Systeme sind die medial-mündlichen Sprachinterfaces der Assistenzsysteme, mit deren Hilfe einfache Funktionen auf

---

1 *massively multi-player online role playing games*

Mobilgeräten gesteuert werden können (*Apples Siri*, *Samsungs Galaxy*). Das Prinzip der virtuellen Assistenz nimmt mit Amazons *Alexa* (*ECHO* und *ECHODot*) nun auch Einzug in unsere Haushalte und Sprachinterfaces in interaktivem Spielzeug (z. B. Mattels *Hello Barbie*) zeichnen kindliche Spielkommunikation auf und sammeln sie als Sprachdaten über das *Internet-of-Things* in einer *Cloud*. Noch wurde die Desktopmetapher zur Steigerung der Usability von digitalen Technologien nicht durch die Assistenzmetapher abgelöst, doch stellt die Sprachsteuerung in bestimmten Anwendungsbereichen eine interessante Alternative dar.

So bedenklich viele der vorstellbaren Szenarien einer Mensch-Maschine-Interaktion pädagogisch, psychologisch und ethisch einerseits sind, so groß scheint andererseits der Wunsch des Menschen a) nach einer intuitiv nutzbaren Schnittstelle zur Technologie und b) nach einem selbst erschaffenen mechanischen Abbild (vgl. Kämer 2008). Im technologieoptimistischen Japan sind anthropomorphe Roboter wie Toshibas *Chihira Kanae* die aktuellste Manifestation dieser Haltung. Technologiehistorisch können menschenähnliche Automatenpuppen aber bereits seit dem hellenistischen Ägypten nachgewiesen werden (vgl. Manhart 2001) und haben geistesgeschichtlich den Diskurs um die Beschaffenheit der menschlichen Physis und Kognition – und damit auch nach deren Einzigartigkeit – immer wieder befeuert (z. B. Vaucansons *Flötenspieler*, Kempelens *Schachtürke* oder die Rechenmaschinen von Leibniz und Pascal). Alan Turing formulierte in dieser Tradition in den 1950er und 1960er Jahren die grundsätzlichen Fragen der Künstlichen-Intelligenz-Forschung (KI), nämlich, a) ob Maschinen denken könnten und b) ob es überhaupt ein Forschungsziel darstellen sollte, Maschinen zum Denken zu bringen.

Der Fokus Turings eigener Bemühungen lag auf der technischen Modellierung menschlicher Kommunikationsfähigkeit, die seit Descartes immer wieder als Alleinstellungsmerkmal des Menschen diskutiert wird. Sein System *ELIZA*, eine virtuelle Psychologin, kann gewissermaßen als „Mutter" aller heutigen Chatbots angesehen werden (vgl. Tewes 2005). Durch einen einfachen *Parsing-Prompting*-Mechanismus mit Schlüsselworterkennung ist des dem System rein deterministisch möglich, mehr oder weniger passend auf Eingaben seiner Userinnen und User zu reagieren (vgl. Kapitel 2). Was modelliert wird, ist also *Dialogperformanz*, mitnichten aber *Dialogkompetenz*, die ein selbstreflexives Bewusstsein der Maschine im Sinne des Locke'schen *Self* voraussetzen würde und sie damit zu intentionalen, zielgerichteten sowie zu freien, spontanen Sprachhandlungen befähigen würde (vgl. Searles Kritik am *Turing-Test*: Gedankenexperiment „Das chinesische Zimmer" von 1980). Menschen dagegen sind sich ihrer Sprachhandlungen grundsätzlich bewusst, auch wenn Bewusstseinsgrade differieren können) und sie folgen in ihrer Kommunikation zwar oft Routinen, aber keinem vorformulierten Programm. Sie sind in der Lage sich kommunikative Ziele selbst zu setzen und diese im Verlauf der Kommunikation dynamisch zu ändern.

Interaktionale Voraussetzungen der Mensch-Maschine-Interaktion

Wenn also ein Mensch mit einem künstlichen Gegenüber in Interaktion tritt, ist das keine Kommunikation unter Gleichen. Vielmehr stellte sich im Rahmen der linguistischen Korpus-Studie, die später detailliert vorgestellt wird, heraus, dass die Mensch-Maschine-Interaktion (*Human Computer Interaction, HCI*) in Abhängigkeit von dialogexternen und dialogimmanenten Faktoren als extrem heterogene Interaktionssituation beschrieben werden muss: Eine menschliche Userin mit Bewusstsein, freiem Willen, Intentionalität und Spontaneität, die als kommunikative Ressource auf ein zeitlebens erworbenes Weltwissen und soziales Wissen (vgl. Habermas 1996) zurückgreifen kann, trifft auf einen Bot als rein deterministischen Mechanismus, dem diese Ressourcen nur äußerst limitiert im Rahmen seines Programms zur Verfügung stehen.

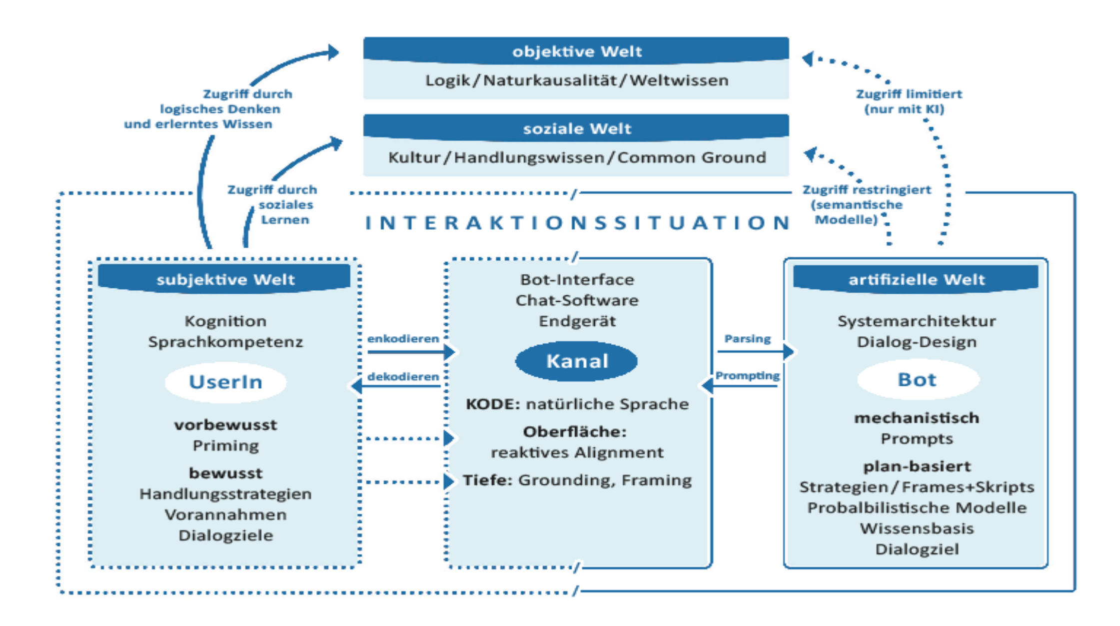


Abbildung 1: Bedingungen der Interaktionssituation HCI: „Vier-Welten-Modell“ (vgl. auch Lotze 2016)

Die Grafik veranschaulicht die Grundbedingungen, von denen Mensch-Maschine-Dialoge beeinflusst werden und die den Interpretationsrahmen für jegliche linguistische Analyse zur HCI liefert. Das Verständnis von subjektiver, objektiver und sozialer Welt als Voraussetzung einer jeden Kommunikation unter Menschen geht auf Habermas‘ *Drei-Welten-Modell* zurück (Habermas 1996) und muss für die HCI um eine vierte Komponente erweitert werden, nämlich um die artifizielle Welt des Bots, die sich durch diverse Restriktionen auszeichnet.

Auch wenn deren Entwicklerinnen und Entwickler selbst die riesige Zahl der internen Kombinationsmöglichkeiten von sprachlichen Bausteinen nicht mehr überblicken, die zu einer Antwort des Bots führen können, bleibt das Grundproblem der fehlenden kommunikativen Ressourcen auch bei innovativen, hochkomplexen Systemen bestehen. Aus linguistischer Perspektive ist dieses Defizit fatal, da Mensch-Mensch-Kommunikation allein auf der Basis geteilten Wissens funktioniert. Menschen machen nicht jede Vorbedingung eines Gesprächsbeitrags und auch nicht jedes Detail in der Kommunikation untereinander explizit, sondern setzen den überwiegenden Teil der in einer Kommunikationssituation theoretisch verfügbaren Informationen als *Common Ground* (vgl. Stalnaker 2002) voraus. Sprachlich formuliert wird nur, was angesichts eines *Partnermodells* des Gegenübers nicht als geteiltes Wissen angenommen und was außerdem als relevant für die Kommunikationssituation eingeschätzt wird vor dem Hintergrund sozialen und kulturellen Wissens. Alles Weitere können die Kommunizierenden im Geiste ergänzen (*Et-Cetra-Prinzip*, Cicourel 1973). Die größte Herausforderung der Mensch-Mensch-Kommunikation ist also das, was implizit bleibt, und auf der Ebene von *Common Ground* und *Partnermodellen* fehlt es allen Dialogsystemen an Wissensressourcen.

Dieses Fehlen geteilter Wissensbestände führt zu logischen Brüchen in der Dialogkohärenz und so oftmals zu absurden Interaktionssituationen, wie dieses Beispiel des eigentlich avancierten Systems *Max* der Universität Bielefeld (Wachsmuth & Kopp 2006) zeigt.

(1) Max: Wo wohnst du? Wie wollen wir weitermachen?
(2) UserIn: in einem land hinter den sieben bergen
(3) Max: Du musst schon genauer erklären, was du meinst Hans.
(Max 1053-1080)

Wer im deutschsprachigen Raum sozialisiert ist, kennt in der Regel die Phrase „hinter den sieben Bergen“ aus dem Märchen *Schneewittchen* der Gebrüder Grimm und kann diese Information als *Common Ground* voraussetzen. Ohne dieses Wissen ist es dem System nicht möglich, die *Ironie* im Beitrag zu erkennen. Eine Rückfrage als Reparatur-Strategie zu bemühen, ist zwar nicht direkt inkohärent, enttarnt das System aber als artifiziellen Gesprächspartner. Den *Turing-Test* (vgl. Artikel Kaerlein) könnte *Max* so nicht bestehen, da er nicht über das Weltwissen verfügt, das im westlichen Kulturkreis als allgemein bekannt gilt.

Systemarchitekturen und Dialog-Designs

Selbstverständlich gehen moderne Systeme heutzutage über Turings einfache Schlüsselworterkennung mit Antwortgenerierung weit hinaus und verfügen in der Regel über eine Reihe an Subsystemen (z.B. zur auf die sprachliche Äußerung abgestimmten Steuerung der Gestik und Mimik eines Avatars o. ä.). Trotzdem findet sich das Grundprinzip auch heute noch in jeder Anwendung. Um Mensch-Maschine-Dialoge mit Fokus auf das

sprachliche Verhalten der UserInnen analysieren zu können, ist es zunächst also von großer Bedeutung ein Grundverständnis für Systemarchitektur und Dialog-Design unterschiedlicher Anwendungen zu haben. Nur so kann z.B. erklärt werden, wodurch Störungen im Dialog hervorgerufen werden. Aus diesem Grund sollen zunächst drei technisch sehr unterschiedliche Herangehensweisen an die Herausforderung HCI skizziert werden.

Die einfache Chatbot-Architektur

In ihrer basalen Form geht die Systemarchitektur eines einfachen Chatbots nicht über einen *Parsing-Prompting*-Mechanismus hinaus; d. h., der Parser analysiert den UserInnen-Input nach *Keywords* und gibt regelbasiert den passenden vorformulierten *Prompt* aus. Dieses Grundprinzip kann seitens des Parsers erweitert werden durch eine *Mustererkennung*, die nicht nur einzelne Wortformen, sondern ganze Wortketten als wiederkehrende *Patterns* im UserInnen-Input erkennt (z. B. „Guten Morgen"). Die Antwortgenerierung wird verbessert durch flexible Antwortstämme (sog. *Templates*) in die schablonenartig in Abhängigkeit von den jeweils eingegebenen *Keywords* und *Patterns* ein passender Baustein eingesetzt und so der System-Output dynamisch generiert wird (z. B. „Mir geht es X (gut, schlecht, so lala etc.)").

Das Grundprinzip der HCI ist insofern störungsanfällig, weil das spontane Verhalten der NutzerInnen bei der Entwicklung des Dialog-Designs nur bedingt antizipiert werden kann.

(1) Max: Ich habe mehrere Lieblingsfilme. Zum Beispiel Terminator (…). Was ist deiner?

(2) UserIn: och, so verschiedene

(3) Max: Dein Lieblingsfilm ist also och so verschiedene?

(Max 1781-1799)

Hätte die Userin in diesem Beispiel einen Filmtitel wie z. B. „Casablanca" genannt, wäre die Antwortgenerierung mit dem *Template* „Dein Lieblingsfilm ist also X" passend gewesen: „Dein Lieblingsfim ist also Casablanca".

Misslingt das *Schlüsselwort-Parsing*, sodass dem *Input-Pattern* kein *Output-Template* zugeordnet werden kann, werden von den meisten Systemen sog. *Default-Antworten* ausgegeben, um den Dialog aufrechtzuerhalten. Diese sind gewissermaßen das Sicherheitsnetz für die weitere Progression. Problematisch an *Default-Antworten* ist, dass sie notwendig Gemeinplätze enthalten, die nicht immer einen thematischen Bezug zum vorangegangenen Input haben (z. B. „Mal so, mal so", „Irgendwas ist immer" usw.).

Schlüsselworterkennung und Antwortgenerierung können erweitert und semantisch strukturiert werden durch sog. *Makros*; d. h. *Synonym*listen, die einer Kategorie zugeordnet werden können.

<category> Verabschiedung

<keywords> bye, cu, cya, exit, quit, ciao, ade, adios, hasta*[2], auf wieder*, tschoe, tschues, tschau, und weg, so long, machs gut, bis bald, bis dann, bis spaeter

Auf diese Weise können also einfache semantische Taxonomien angelegt werden. Das gilt auch für (*Hypero- / Hyponym*-Relationen wie *„[Tier] > [Haustier] > [Katze, Hund, Chinchilla usw.]“*. Dass eine semantische Analyse durchgeführt wird, bedeutet auf dieser Ebene aber nicht, dass der Parser die Satzaussage verarbeiten könnte und schon gar nicht, dass das System etwas „versteht“. Denn die Fähigkeit „zu verstehen“, setzt ein Bewusstsein voraus, das sich selbst erkennt und in Relation zum Verstandenen setzen kann (vgl. Begriff der „Intentionalität“, Searle 2006, S. 171ff).

Von besonderer Relevanz für störungsfreie Dialoge auf der Basis der einfachen Chatbot-Architektur sind extrem umfangreiche, semantisch gut strukturierte Schlüsselwortlexika und zehntausende hinterlegte Patterns und Templates. Auf diese Weise funktioniert die Interaktion mit einfachen Chatbots relativ frei und flexibel, dadurch werden aber Dialogverläufe weniger antizipierbar und die Systeme damit störungsanfälliger. Die Dialoge verlaufen dadurch sprunghaft und inkohärent. Einfache Chatbots mit völlig freier UserInnen-Führung eignen sich deswegen vor allem für den Entertainment-Bereich.

Plan-basierte Systeme und Dialog-Skripte

Ein Versuch Dialog-Designs so zu gestalten, dass Dialogverläufe antizipierbarer werden, wird bei sog. *plan-basierten Systemen* unternommen. Diese arbeiten vordefinierte Handlungspläne Schritt für Schritt ab. Dabei sind die Handlungsoptionen in jedem Arbeitsschritt durch Entscheidungsbäume limitiert. Wie die Systemarchitektur folgt auch das Dialog-Design einem vordefinierten Plan, also einem *Dialogskript*. Solche Systeme eignen sich als Assistenzsysteme für klar umrissene Aufgabenbereiche wie z.B. einem Verkaufsgespräch oder der Ausgabe von Informationen zu Exponaten in einem Museum (vgl. System *Max* im *Heinz Nixdorf Museum*).

Aus linguistischer Perspektive sind Gespräche unter Menschen trotz der theoretisch gegebenen Freiheit nicht völlig unstrukturiert, sondern folgen im Alltag auch oft Mustern und Routinen (vgl. Sacks 1992). Vor allem in immer wiederkehrenden Interaktionssituationen können sich konventionalisierte Skripte herausbilden (vgl. z. B. für Verkaufsgespräche Schlobinski 1982, 1997).

(1) P: ein kopfsalat

(2) M: ein kopfsalat. (4.0) so (.) außerdem?

(3) P: drei zitronen (Schlobinski 1997, S. 13)

---

[2] Der Kleene-Operator fungiert hier als Platzhalter für verschiedene Varianten (*hasta luego, hasta mañana, hasta la vista*).

Der Kunde P benennt das gewünschte Produkt. Darauf folgt eine Bestätigung des Verkäufers M und die ritualisierte Nachfrage „außerdem?“. Diese konventionellen Alltags-Skripte werden in plan-basierten Dialogagenten für die HCI nutzbar gemacht. Diese enge Form der UserInnen-Führung ist weniger störungsanfällig als einfaches Parsing und Prompting, aber wesentlich unflexibler.

KI und „Big-Data-Analysis“

Für die Konstruktion der Systemarchitekturen innovativer Systeme wie z.B. *IBMs Watson* wird auf moderne Methoden der „Big-Data-Analysis“ und / oder der klassischen KI-Forschung zurückgegriffen. Die derzeit besten Dialogagenten vereinen alle grundverschiedenen Herangehensweisen als hybride Systeme mit einfacher Chatbot-Architektur unter Rückgriff auf große Datenbanken und klassische KI-Kalküle.

In einer „Big-Data-Architektur“ wird auf Dialog-Datenbanken zurückgegriffen, die Milliarden von aufgezeichneten Mensch-Mensch-Konversationen oder früheren Mensch-Maschine-Interaktionen enthalten. Über ein statistisches Modell wird für jeden Dialogschritt (*Turn*) die jeweils wahrscheinlichste Reaktion aus der Datensammlung ausgewählt. In der simpelsten Form heißt das: Der Bot recycelt die gespeicherten Beiträge seiner früheren UserInnen und gibt sie in einer passenden Dialogsituation als seine eigenen wieder aus (vgl. u. a. Carpenters *Cleverbot*).

Problematisch an dieser Herangehensweise ist, dass Dialoge nicht immer gleich verlaufen und es daher alles andere als trivial ist (wenn nicht sogar unmöglich!), die Relevanz eines hinterlegten Beitrags für die aktuelle Dialogsituation rein statistisch zu ermitteln. Diesem Problem soll mittels einer durch klassische KI gestützten Analyse der Propositionen und ihrer logischen Konnektoren in (und zwischen) den jeweiligen Dialogbeiträgen entgegengewirkt werden. Sie sollte es den Systemen eigentlich ermöglichen, auch inferenzielle Schlüsse aus formulierten Bedingungen zu ziehen. Leider ist es aber kaum möglich, logische Kalküle über natürlich-sprachliche Aussagen im Dialog zu rechnen, weil diese inhaltlich so viel voraussetzen (vgl. *Et-Cetera-Prinzip* s.o.).

Ein zweites Problem stellt die Inkonsistenz der Ressource „Big-Data“ dar. Wenn unterschiedliche UserInnen ihre jeweiligen persönlichen Daten und sehr unterschiedlichen Überzeugungen kommunizieren, kann man aus diesem Patchwork-Input nicht den Output für eine konsistente Bot-Persona ableiten. So kann *Cleverbot* z.B. mehrfach im Chat die Gender-Rolle wechseln, je nachdem, ob Userinnen oder User-Beiträge recycelt werden (vgl. *automatisches Nexting* Christian 2012, S. 25). Noch problematischer als Ressource werden die Beiträge früherer UserInnen, wenn sich unter diesen *Trolls* oder Agitatoren befinden, wie bei einem der ersten Social Bots 2016 geschehen. Der Twitter-Bot *Tay*, der wahllos in Retweets weiterleitete, was UserInnen eingaben, musste vom Netz genommen werden, als er plötzlich faschistische Parolen postete.



Abbildung 2: Rechtsextreme Retweets des Twitter-Bots Tay

Seitens der Funktionsweise der Systeme liegen aus (computer-)linguistischer Perspektive die größten Herausforderungen der HCI-Forschung auch heute noch in der Erstellung logisch kohärenter Dialoge. Dies funktioniert nur durch eine zuverlässige semantische Analyse der Einzelbeiträge und deren kohärenter Abfolge über längere Dialogsequenzen unter Berücksichtigung ihrer sozialen und situativen Angemessenheit und des vorausgesetzten geteilten Wissens der Interagierenden. Sind diese Bedingungen nicht gegeben, sind die Dialoge nicht nur störungsanfällig, sondern wirken in jeder Hinsicht sprunghaft und unlogisch bis paradox und absurd.

Sprachliche Merkmale der HCI: Ergebnisse der linguistischen Analyse

Wie sich Menschen verhalten, die sich mit einer solchen Technologie konfrontiert sehen, war Untersuchungsgegenstand einer mikro-diachron (2000-2012) angelegten linguistischen Korpusstudie zu unterschiedlich avancierten Chatbots und den sprachlichen Strategien ihrer UserInnen (Lotze 2016)[3]. Dabei stand die Frage im Zentrum, inwiefern Menschen typische sprachliche Verhaltensweisen aus der Mensch-Mensch-Kommunikation auf die HCI übertragen, z.B. in Form eines interaktiven Alignments (Pickering & Garrod 2004) oder ob sie neue Interaktionsformen mit der Maschine finden im Sinne eines sog. „Computer-Talk“ (Zoeppritz 1985; Krause et al. 1992; Fischer 2006, 2016).

[3] Dissertation an der Leibniz Universität Hannover unter Betreuung von Prof. Dr. Peter Schlobinski und Prof. Dr. Kerstin Fischer (Sonderburg, DK), Publikation: Lotze (2016). *Chatbots – eine linguistische Analyse*. Lang.

Linguistische Theorien zur Analyse von HCI

HCI als in ihren Vorbedingungen für Bot und Mensch äußerst heterogene Interaktionsform zeichnet sich durch eine Reihe an interaktionalen Problemen aus, die sich auf unterschiedlichen linguistischen Ebenen manifestieren, wie am Beispiel des Systems *Elbot* (Roberts 2006) gezeigt werden kann.

(1) Elbot: Vielleicht liegt das an mehr Erfahrung.

(2) UserIn: vielleicht auch an mangelnder zahl von bits und bytes

(3) Elbot: Das Leben als Mensch soll ja nicht so einfach sein.

(4) UserIn: floskelarsch

(Elbot 2963-2996)

Das erste interaktionale Problem (oder Potenzial?) manifestiert sich in Form von *interaktivem Alignment* (Pickering & Garrod 2004). Der/die UserIn übernimmt lexikalische und syntaktische Strukturen aus Elbots Gesprächs-Turn (*vielleicht, an* + Nominalphrase). Die *Konstruktionsübernahme* ist ein hochfrequentes Verfahren in der Mensch-Mensch-Kommunikation und dient u.a. auch der Etablierung von *Common Ground*. Eigentlich müsste das System den Beitrag des/r UserIn durch diese Wiederaufnahme *(Rekurrenz)* von zuvor selbst Ausgegebenem zuverlässiger parsen können. Dies ist hier leider nicht der Fall, da der Parser mit einem zweiten hochfrequenten Phänomen aus der Mensch-Mensch-Kommunikation kämpft, nämlich der *Adjazenzellipse*. Der/die UserIn gestaltet den Chatbeitrag nicht in Form eines ganzen Satzes, sondern lässt das bereits bekannte *„liegt das"* aus und setzt es als geteiltes Wissen voraus. Dieser nicht satzwertige, also *elliptische*, Anschluss kann vom System nicht logisch rückbezogen werden. Um die Störung zu „tarnen" greift als System-Reaktion das Sicherheitsnetz der *Default-Antworten*: *„Das Leben als Mensch soll ja nicht so einfach sein;"* ist ein Gemeinplatz, der von Elbot öfter ausgegeben wird und der manchmal mehr und manchmal weniger kohärent anschließt. Diese Pseudo-Kohärenz, die sich im Rahmen der Studie als typisch für die HCI herausgestellt hat, wird von dem/r UserIn direkt durch einen ausgesprochen treffend gewählten *Vulgarismus* (im Netz auch *Flaming*) sanktioniert. Beschimpfungen des Systems seitens der UserInnen sind in der chat-basierten HCI hochfrequent. Dass das System nicht zur logisch-kohärenten Verarbeitung der UserInnen-Eingabe in der Lage ist, sollen *Kohäsionsmarker* im Output des Bots verschleiern, die als verweisende Elemente an der Textoberfläche auf einen logischen Aufbau und einen roten Faden im Gespräch schließen lassen (z.B. rückverweisendes *das* im ersten Turn).

Als Fazit aus diesem Textbeispiel kann festgehalten werden, dass UserInnen einerseits Muster aus der Mensch-Mensch-Kommunikation zu übertragen scheinen und andererseits dem Bot gegenüber Äußerungen tätigen, die man in Dialogen unter Menschen extrem selten findet (vgl. Vulgarismus *floskelarsch*). Um die Sprache der UserInnen besser zu verstehen, ist es also wichtig, sie aus unterschiedlichen Perspektiven und auf unterschied-

lichen Ebenen zu untersuchen, die repräsentiert werden durch die verschiedenen theoretischen Zugänge, die in der Studie verfolgt wurden (vgl. Lotze 2016).

- Die **Konversationsanalyse** in ihren Grundzügen nach Harvey Sacks in Bezug auf basale Interaktionsmuster wie u.a. *Adjazenzstrukturen*. Hier sind vor allem die tiefe Einsicht für die HCI relevant, dass menschliche Kommunikation nicht willkürlich und chaotisch verläuft, sondern von einer „Order at all points" ausgegangen wird, die definiert werden kann. Daraus leitet sich auch das Prinzip der „bedingten Erwartbarkeit" von Folgebeiträgen ab, was fruchtbar für die HCI gemacht werden kann, wo UserInnen-Beiträge antizipierbar sind.
- Das **interaktive Alignment-Modell** nach Pickering und Garrod als *psycholinguistischer* Versuch, persistente Strukturen, also lexikalische und syntaktische Konstruktionsübernahmen, im Dialog auf kognitiver Ebene zu erklären. Für die HCI besonders interessant ist, dass Pickering und Garrod annehmen, dass dieses Verhalten auch in der Mensch-Mensch-Kommunikation ohne höhere Kognition rein mechanistisch ablaufe. Diese Position wurde allerdings mehrfach scharf kritisiert (u.a. Schegloff 2004).
- **Kohärenzstrukturen** als logisches Skelett von Dialogen und Marker für diesen roten Faden in Form von **Kohäsion** an der Textoberfläche lassen sich nach Kriterien der *Textlinguistik* beschreiben (Givón 1983, Tidge 1997, Linke et al. 2004, Brinker 2010).
- Seit den 1980er Jahren wird die Möglichkeit eines „Computer-Talk" als besonders simple oder gerade sehr explizite Sprache diskutiert, die UserInnen nur artifiziellen Dialogsystemen gegenüber verwenden (Zoeppritz 1985; Krause et al. 1992; Fischer 2006, 2016).

Es handelt sich also um ein über die Teildisziplinen der Linguistik hinaus inter- oder transdisziplinäres Projekt zwischen Linguistik, Psychologie, Soziologie, Philosophie und Informatik.

Datenbasis und Methode

Im Rahmen der linguistischen Korpusanalyse wurden vier Systemkorpora zur HCI erhoben und ein Vergleichskorpus zur chat-basierten Mensch-Mensch-Kommunikation (*CMC*, aus: *Dortmunder Chat-Korpus*, Storrer, Beißwenger 2004). Um vergleichbare Szenarien untersuchen zu können, wurden ausschließlich Infobots[4] ausgewählt sowie der Online-

[4] *Twipsy* (2000): Kiwilogic (Wagner, Schmitt), *Karlbot* (2004): Kiwilogic (Prehn, Sharp, Gövert), *Max* (2008): Uni Bielefeld (Wachsmuth, Kopp), *Elbot* (2012): Artificial Solutions (Roberts) (als Infobot auf freenet.de)

Help-Desk der Universitätsbibliothek Dortmund (Bib) für die Mensch-Mensch-Kommunikation.

Insgesamt wurden 183 Dialoge feinkörnig analysiert. Dabei wurden quantitative und qualitative Verfahren kombiniert. Die quantitative Analyse erfolgte als relative Häufigkeiten der Tokens pro Teilkorpus. So konnten linguistische Vergleiche auf unterschiedlichen Ebenen vorgenommen werden, nämlich zwischen UserIn und System, HCI und CMC (unter Menschen) sowie ein diachroner Vergleich zwischen den einzelnen Systemen.

Für das Phänomen des interaktiven Alignments erlaubte die hohe Anzahl der Belege eine Auswertung Dialog für Dialog und damit auch eine inferenzstatistische Auswertung. Außerdem wurde die Distanz zwischen persistenten Strukturen (Prime und Persistenz) gemessen. Begleitet wurde die quantitative Auswertung durch qualitative Untersuchungen von Ausschnitten aus den Korpora, was gerade im Bereich Kohärenz und Kohäsion unerlässlich war (z.B. bei Störungen und Reparaturen).

Ergebnisse der linguistischen Analyse

In der Studie stellten sich sprachliche NutzerInnen-Strategien als äußerst vielseitig heraus. Denn einerseits wird ein grundsätzliches Verständnis von interaktivem Austausch auf die HCI übertragen, mit dem auch unterschiedliche interessante linguistische Phänomene aus der Mensch-Mensch-Kommunikation auf die HCI transferiert werden. Andererseits finden sich auch gegenüber den moderneren Bots noch Anzeichen eines sog. „Computer-Talks“, der darauf schließen lässt, dass sich die NutzerInnen eben gerade nicht der Illusion einer natürlichen Kommunikation hingeben, sondern mit knappen Befehlen eine Maschine bedienen. Zwischen diesen Extrempunkten bewegt sich HCI, und ein und der/die selbe NutzerIn kann in verschieden Dialogsequenzen durchaus beide Herangehensweisen ausprobieren. Beispielhaft sollen im Folgenden ausgewählte Ergebnisse diskutiert werden, an denen man die Grundtendenzen der chatbasierten HCI gut exemplifizieren kann.

Dialogische Basisstrukturen: Die qualitative Analyse nach Parametern der Konversationsanalyse hat ergeben, dass die UserInnen durchaus Strukturen verwenden, die der Kommunikation unter Menschen gleichen: u.a. wie im obigen Beispiel *Adjazenzellipsen* in großer Zahl. Diese werden auch dann gebraucht, wenn das System i.d.R. ganze Sätze ausgibt. Die von den UserInnen bemühten grundsätzlichen Dialogstrukturen wie Adjazenzpaare oder Gesprächsrahmenstrukturen, also ritualisierte Begrüßungen und Verabschiedungen u.a., fallen in der HCI quantitativ immer hinter die Vergleichswerte aus der CMC zurück. Besonders bei älteren Systemen finden sich asymmetrische Begrüßungen und vor allem asymmetrische Verabschiedungen oder unvermittelte Gesprächsabbrüche. Auch allgemein lässt sich das UserInnen-Verhalten meist aus dem Dialog-Design des Systems erklären. Ein bestimmtes System fördert in einer bestimmten Dialogphase aufgrund von Design-Entscheidungen ein bestimmtes Verhalten seitens der UserInnen.

Im Vergleich zwischen Mensch-Maschine- und Mensch-Mensch-Kommunikation lassen sich größere Abweichungen im Sprachgebrauch zwischen UserIn und System in der HCI als zwischen den UserInnen in der CMC feststellen. Z.B. kann eine Divergenz zwischen extrem kurzen Dialogbeiträgen (*Turns*) der UserInnen und extrem langen System-Turns beobachtet werden. So sind die Turns der UserInnen in den untersuchten HCI-Chats im Durchschnitt nur drei Wortformen lang, während die des Systems über 20 Wortformen verfügen. Dies bedeutet für die chatbasierte HCI noch eine Zuspitzung des Befunds von Fischer für die mündliche Mensch-Roboter-Interaktion (HRI), die eine durchschnittliche Turn-Länge von fünf Wortformen (5 WF) nachweist (Fischer 2006, 2010, 2016).

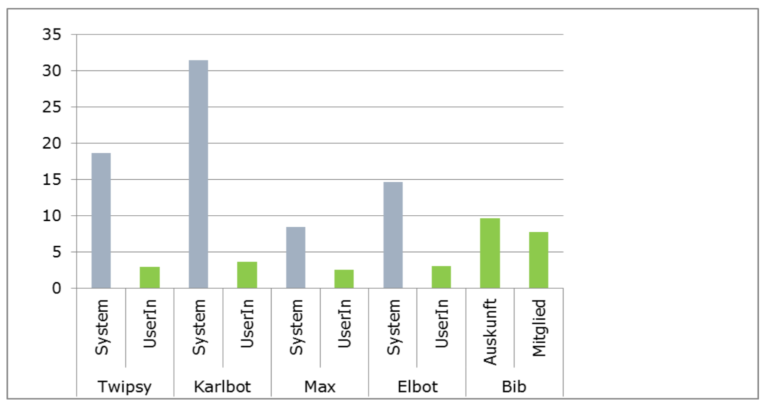


Abbildung 3: ***Vergleich UserIn/System:*** *Divergenz bei Turn-Längen (Anzahl der Wortformen)*

UserInnen in HCI und HRI sind also äußerst einsilbig, besonders in der chatbasierten Mensch-Maschine-Interaktion, bedingt durch die Eingabemodalitäten. Während die Bot-Turns i.d.R. *konzeptionell monologisch* als in sich geschlossene Einheiten gestaltet werden, scheinen die UserInnen an einem *dialogischen Konzept* festzuhalten. Das äußert sich z.B. in einer großen Zahl an *Adjazenzellipsen*. Gerade die älteren Systeme verfügen über vorformulierte Prompts, die den Eindruck erwecken, als seien sie aus einer Informationsbroschüre herauskopiert worden. Man könnte also ein Kontinuum der Dialogizität ansetzen, auf dem sich die Beiträge der UserInnen und die der Bots an unterschiedlichen Enden des Spektrums einordnen ließen. Diese Grundtendenz hat Implikationen für den Gebrauch einer Reihe an linguistischen Formen wie *Modalpartikeln* oder *Kohäsionsmittel* (s. detaillierter Lotze 2016).

Interaktives Alignment: Unter interaktivem Alignment versteht man in der Psycholinguistik die menschliche Tendenz, sich einem Gegenüber im Gespräch sprachlich anzupassen. Wir neigen dazu, einzelne Wortformen oder komplexere Ausdrücke von unserem/r GesprächspartnerIn zu übernehmen. So etablieren Menschen im Verlauf der Kommunikation

dynamisch ein gemeinsames Dialoglexikon als eine Form von *implizitem Common Ground* (vgl. Pickering & Garrod 2004). Dabei handelt es sich um ein ganz natürliches, hochfrequentes Dialogphänomen. Im psycholinguistischen Fachdiskurs wird diskutiert, inwiefern den sprachlichen *Rekurrenzen* bzw. *Persistenzen* ein vorbewusstes Verhalten oder eine bewusste Strategie zu Grunde liegen.

Für die HCI konnten im Rahmen der Studie sowohl lexikalische als auch syntaktische Konstruktionsübernahmen seitens der UserInnen aus dem System-Output beobachtet werden. Diese waren allerdings seltener als im Vergleichskorpus zur Mensch-Mensch-Kommunikation. Dennoch können sie vor allem bei den UserInnen der neueren Systeme als Indikator für eine Übertragung des grundsätzlichen Interaktionskonzepts aus der Mensch-Mensch-Kommunikation angesehen werden.

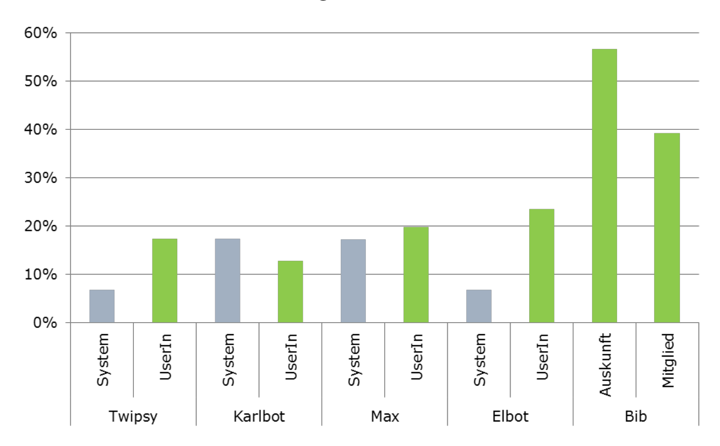


Abbildung 4: ***Vergleich UserIn/System:*** *lexikalische Persistenzen (pro Turn-Konstruktionseinheit)*

Je besser das System funktioniert, desto mehr lassen sich NutzerInnen auf die Illusion einer natürlichsprachlichen Interaktion ein und desto mehr Konstruktionsübernahmen finden sich in den entsprechenden UserInnen-Korpora. Bei den Konstruktionsübernahmen durch das System handelt es sich um Artefakte, die durch die Keyword-Analyse und Antwort-Generierung entstehen (*Pseudo-Alignment*, vgl. Lotze 2016).

Die UserInnen der älteren Systeme nutzen *Persistenzen* teilweise bewusst als *Reparatur-Strategie*, weil sie davon ausgehen, dass der Parser Keywords analysieren kann, die das System zuvor selbst ausgegeben hat. Das ist aber nicht immer der Fall, da Parsing und Prompting separate Prozesse darstellen. Dass ein spezieller Alignment-Generator (z.B. SPUD*Prime* (Buschmeier, Kopp und Bergmann (2010)), der interaktives Alignment seitens des Systems strukturiert simuliert, einen positiven Einfluss auf HCI-Dialoge hätte, kann nur vermutet werden.

Insgesamt lässt sich aber konstatieren, dass vieles darauf hinweist, dass die kognitive Dialogverarbeitung der UserInnen in der HCI nicht grundsätzlich anders abläuft als in der CMC (differenzierter dazu: Lotze 2016). So folgt z.B. die Zerfallsrate der Primes, die als Persistenzen wiederaufgenommen werden, der *„Vergessenskurve“* nach Ebbinghaus; d.h., für UserInnen in der HCI wird eine vom System ausgegebene sprachliche Struktur in ähnlicher Weise wie in der Mensch-Mensch-Kommunikation nach und nach weniger relevant.

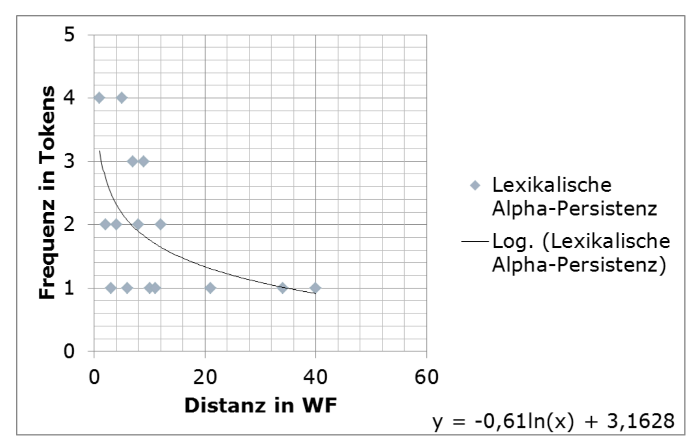


Abbildung 5: ***Distanz/Frequenz-Analyse:*** *Zerfallsrate der Primes als logarithmische Funktion (vgl. Ebbinghaus (1985), „Vergessenskurve“)*

Dialogkohärenz: Wie eingangs erläutert, liegt die größte Herausforderung für Dialogsysteme in der Generierung logisch kohärenter Anschluss-Turns, die inhaltlich auf dem UserInnen-Input aufbauen. Längere Bot-Beiträge sind zwar immanent kohärent und erwecken daher den Eindruck, der Bot habe ein Konzept von logisch-semantischer Kohärenz, aber am Turn-Wechsel wird deutlich, dass es den Systemen nicht immer gelingt *logisch kohärent* und schon gar nicht *sprachlich kohäsiv* an einen UserInnen-Turn anzuschließen. Die Systeme haben also Probleme mit turn-übergreifender Referenz und sequenz-übergreifende Bezüge werden gar nicht hergestellt. Und auch einige gelungene Beispiele, die die Illusion von kohärenten Anschlüssen suggerieren, können nicht darüber hinwegtäuschen, wenn Defizite des semantischen Parsings zu Störungen in der Kohärenzstruktur und zu logischen Brüchen führen.

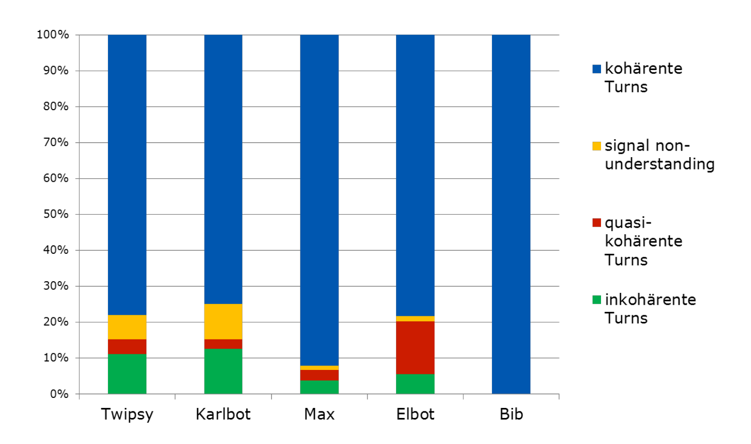


Abbildung 6: ***Vergleich zwischen den Systemen:*** *Kohärenz der Turn-Anschlüsse der UserInnen in Prozent (pro Teil-Korpus)*

Zwar schließen nur 10% der Turns wirklich inkohärent an, aber zusammen mit *Default-Antworten* (s.o.) als eben nur *quasi-kohärenten* Turns und offenen Fehlermeldungen und Rückfragen (*signal non-understanding*) ergeben sich 20% problematische Anschlüsse, die zu Störungen des Dialogs führen. Nur 80% kohärente Turn-Anschlüsse reichen nicht aus, um die Illusion eines natürlichen Dialogs zu schaffen, denn solche grundsätzlichen logischen Brüche kommen im Vergleichskorpus zu CMC (Mensch-Mensch: s. *Bib*) nicht vor.

Und so zeichnet sich auch ein Zusammenhang zwischen logischen Brüchen und (entnervten) Gesprächsabbrüchen durch die UserInnen ab.

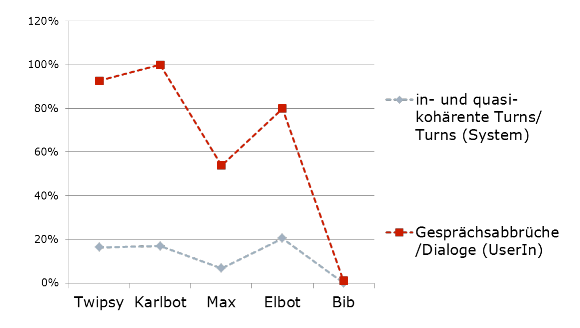


Abbildung 7: Kohärenz / Gesprächsabbrüche in Prozent (pro Dialog)

Auf dieser Basis muss die Frage diskutiert werden, wieviel Freiheit die UserInnen-Führung verträgt. Denn der Dialogagent Max schneidet durch seine hinterlegten Skripte in diesem Bereich als einziges System besser ab.

„Computer-Talk“: Im Rahmen der Studie können einige sprachliche Phänomene seitens der UserInnen nachgewiesen werden, die in dieser Form nur einer Maschine gegenüber geäußert werden können. Es handelt sich also um Belege für eine neue Online-Variante des „Computer-Talks“, die stark abweicht von Zoeppritz’ ursprünglicher Konzeption als einfache und / oder explizite Befehlssprache (Zoeppritz 1989, S. 117).

Im Gegensatz zu Fischers (2006) „speakers attend to politeness“ in der mündlichen HRI finden sich in den medial schriftlichen Korpora bis zu sechs Belege für sprachlich eindeutig unhöfliches Verhalten der UserInnen pro Dialog meist in Form von *Flaming*, obwohl beim Dialog-Design der Systeme auf einfache Höflichkeitsmarker wie *danke*, *bitte* oder *Entschuldigung* geachtet wurde (vgl. Brown & Levinson 1987). Dies hängt sicher mit dem Medium Chat zusammen und kann als Argument für „Computer-Talk“ angesehen werden. Denn in der Mensch-Mensch-Kommunikation würde ein derart vulgäres Verhalten stärker sanktioniert und ginge einher mit einem sozialen Gesichtsverlust. Aus diesem Grund findet es sich auch nicht im CMC-Vergleichskorpus. Vergleicht man die unterschiedlichen Systeme untereinander, wird aber auch die Kontextabhängigkeit dieses UserInnenverhaltens deutlich. UserInnen beschimpfen das System einerseits als Reaktion auf Störungen, andererseits aber auch, um die Grenzen des Dialog-Designs zu testen (wie u. a. auch durch *Anspielungen, Ironie, Fremdsprachen* etc.).

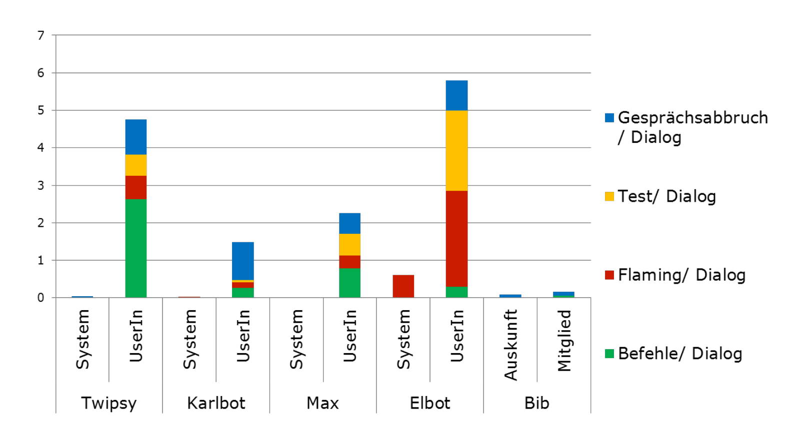


Abbildung 8: ***Ergebnisse „Computer-Talk“:*** *Sprachliche Marker für Unhöflichkeit (pro Turn-Konstruktionseinheit)*

Bei älteren Systemen ist die strukturelle Divergenz zwischen Bot und Mensch größer als bei den neueren. Mögliche Gründe dafür könnten sein, dass die System-Turns bei den neueren Systemen dialogischer gestaltet sind und so weniger „Computer-Talk“ evozieren. Außerdem hat sich die Rezeptionshaltung der modernen UserInnen verändert, sodass sich

in der heutigen chatbasierten HCI nur noch zwei Formen von „Computer-Talk“ finden (vgl. Lotze 2016, S. 369):

- Proaktiver „Computer-Talk“ im Sinne von Tests und *Flaming*
- Reaktiver „Computer-Talk“ im Sinne von „natürlichsprachlichem Weiterklicken“ (*ok, ja, weiter*), bei dem sich NutzerInnen passiv vom Bot durch Dialogskripte führen lassen und jeden Dialogschritt durch ihre Eingaben bestätigen als natürlichsprachliches Äquivalent zur Menü-Navigation im *Graphical User Interface (GUI)*

Fazit: HCI als neue sozio-linguistische Praktik

Durch Zusammenführung der Ergebnisse aus den jeweiligen linguistischen Teilbereichen lässt sich ein konkreteres Bild der heterogenen Interaktionsform HCI zeichnen, die in Abhängigkeit von externen und internen Faktoren in ihrer sprachlichen Ausprägung stark variieren kann.

Systemarchitektur	UserInnen-Führung: geführt – frei – hybrid
Dialog-Design	• Monolog – Dialog • Schriftlichkeit – Mündlichkeit • Dialogphasen • Umgang mit Störungen
UserIn	• UserInnen-Typ • bewusst strategisch - vorbewusst oder routiniert

Abbildung 9: ***Charakteristika der Human-Computer-Interaktion****: HCI als heterogenes Konstrukt in Abhängigkeit von diversen Faktoren*

Durch die Tabelle der dialog-immanent relevanten Faktoren in Abbildung 9 kann nun der Platzhalter „Dialog-Design“ aus dem Modell zu den dialogexternen Faktoren und Grundbedingungen der HCI aus Abbildung 1 mit Inhalt gefüllt werden. Denn in dieser Form können die Wechselwirkungen von Dialog-Design und (sozio-)linguistischen Praktiken der UserInnen spezifiziert werden. Dazu wird angenommen, dass sich HCI-Beiträge innerhalb von Kontinua bewegen:

- Zwischen Monolog und Dialog
- Zwischen konzeptioneller Schriftlichkeit und Mündlichkeit
- Und damit auch zwischen sozialer Nähe und Distanz

Die UserInnenführung ist entweder eng geführt wie eine Menü-Navigation, frei oder hybrid und wechselt passagenweise z.B. mit unterschiedlichen technischen Lösungen für un-

terschiedliche Dialogphasen. Das führt dann auch zu passagenweisen Veränderungen im UserInnenverhalten und die UserInnensprache verändert sich innerhalb der Kontinua. Hierin liegt ein wichtiger Aspekt der dialog-immanenten und userInnen-immanenten Heterogenität der HCI. Es gibt also aus linguistischer Perspektive nicht die eine HCI, sondern eine Vielzahl unterschiedlicher Ausprägungen in Abhängigkeit von den o.g. Faktoren.

Die Asymmetrie zwischen UserIn und System lässt sich auch in den in der Tabelle angeführten Spektren fassen. So orientieren sich die UserInnen der älteren Systeme sprachlich eher am dialogischen Pol des Spektrums, während die System-Beiträge klar monologisch gestaltet sind. Bei den neueren Systemen liegen Mensch und Maschine innerhalb der Kontinua näher beieinander als bei den älteren Systemen, sodass eine *Konvergenz/Divergenz-Hypothese* aufgestellt werden kann: „Je besser das jeweilige System funktioniert, desto geringer fällt die Divergenz zwischen UserInnen- und Systemsprache aus. Dabei nähern sich die Werte der Systeme und die der UserInnen einander gleichermaßen an“ (Lotze 2016: 360f). Daraus folgt als Prognose eine weitere beiderseitige Annäherung in der Zukunft, d.h., wir passen uns nicht wie in den 1980ern erwartet vollständig den Systemen an, aber die Systeme werden sich auch nicht vollständig den Menschen anpassen. Mensch-Maschine- und Mensch-Mensch-Dialoge werden sich in näherer Zukunft nicht bis zur Übereinstimmung annähern, da die Restriktionen von künstlicher Intelligenz, nämlich das Fehlen von *Bewusstsein, semantisch-strukturiertem Weltwissen, sozialem Wissen, Partner-Modell und Common Ground, Intentionalität* sowie *avanciertem inferenziellem Schließen*, eine technische Realisierung menschlicher Dialogfähigkeit unmöglich machen. Wahrscheinlicher ist, dass Menschen eine neue (sozio-)linguistische Praktik erlernen, nämlich die jeweils anwendungsadäquate Interaktionsform mit unterschiedlichen Systemen wie Social Bots, Assistenzsystemen, Robotern usw., die zwar intuitiv und natürlichsprachlich erfolgt, nicht aber menschliche Kommunikation abbildet.

Aus linguistischer Perspektive problematisch ist vor allem die Illusion eines menschenähnlichen Gegenübers, das dann aber weit hinter die menschliche Sprachkompetenz zurückfällt. Je nach NutzerInnen-Typ lassen sich Menschen mehr oder weniger auf die Illusion HCI ein (vgl. Fischer 2006). Obwohl die Dialoge immer noch sprunghaft und inkohärent sind, zeigen NutzerInnen nämlich teilweise auch die Tendenz, sich dem Sprachgebrauch von Systemen anzupassen (vgl. Alignment). In dieser Gemengelage steckt die Gefahr, dass auch das sprunghafte, unlogische und inkonsistente Dialogverhalten des Bots von (besonders jüngeren) UserInnen kopiert werden könnte. Welche sprachlichen, kommunikativen und sozialen Auswirkungen durch einen solchen Zusammenhang bedingt würden in unterschiedlichen Szenarien wie z.B. der kindlichen Interaktion mit dialogfähigem Spielzeug oder dem Lernen mit interaktivem Tutorensystem, ist derzeit noch völlig unabsehbar und stellt ein extrem zukunftsrelevantes, interdisziplinäres Forschungsdesiderat dar – neben der Linguistik auch für Informatik, Psychologie, Pädagogik, Ethik oder Rechtswissenschaften. Somit muss man die Relevanz der frühzeitigen Forschung zur HCI/HRI betonen: Da wir in immer mehr Lebensbereichen mit Dialogsystemen in Interaktion treten werden,

wird es vermehrt wichtig, Strategien im Umgang mit Möglichkeiten und Grenzen der Systeme zu entwickeln. Und dies ist nur möglich, wenn wir im ersten Schritt die Zusammenhänge und Mechanismen der HCI in ihrer ganzen Heterogenität und anwendungsspezifischen Vielfalt verstehen lernen (*Medienkompetenz*). Erst im zweiten Schritt können wir so mit innovativen Anwendungen zu einem homogeneren Gesamtbild der HCI gelangen (*technologischer Fortschritt*). Und im dritten Schritt müssen wir uns als Menschen wieder von der Technologie emanzipieren, indem wir ihre Grenzen akzeptieren und mit der uns eigenen Kreativität neue (sozio-)linguistische Praktiken zur Nutzbarmachung der Systeme im Alltag (*digitale Aufklärung*) entwickeln.

Literatur

Brinker, Klaus (2010): *Linguistische Textanalyse.* Berlin: Schmidt.

Brown, Penelope & Levinson, Stephan C. (1987): *Politeness. Some Universals in Language Usage*. 2. Aufl. Cambridge: Cambridge University Press.

Buschmeier, Hendrik; Kopp, Stefan & Bergmann, Kirsten (2010): Modelling and Evaluation of Lexical and Syntactic Alignment with a Priming-Based Microplanner. Online verfügbar unter: http://dbonline.igroupnet.com/ACM.FT/1890000/1880377/p85-buschmeier.pdf.

Christian, Brian (2012): *The most human Human – What Artificial Intelligence Teaches Us About Being Alive*. Bottrop: Anchor.

Cicourel, Aaron Victor (1973): Basisregeln und normative Regeln im Prozess des Aushandelns von Status und Rolle. In: Arbeitsgruppe Bielefelder Soziologen (Hrsg.): *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit.* Frankfurt a. M.: Fischer Taschenbuch Verlag, 1, S. 147–188.

Ebbinghaus, Hermann (1985): *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie.* Leipzig: Duncker & Humblot.

Fischer, Kerstin (2006): *What Computer Talk is and Isn't: Human–Computer Conversation as Intercultural Communication*. Saarbrücken: AQ.

Fischer, Kerstin (2010): Why it is interesting to investigate how people talk to computers and robots. Introduction to the special issue. *Journal of Pragmatics,* 42, S. 2349–2354.

Fischer, Kerstin (2016): *Designing Speech for a Recipient. The Role of the Addressee in So-called Simplified Registers*. Amsterdam: John Benjamins.

Givón, Talmy (1983): *Topic continuity in discourse – A quantitative cross-language study*. Amsterdam: Benjamins, S. 5–38.

Habermas, Jürgen (1993): *Theorie des kommunikativen Handelns*. 2 Bände. Frankfurt a. M.: Suhrkamp.

Krämer, Nicole C. (2008): *Soziale Wirkungen virtueller Helfer. Gestaltung und Evaluation von Mensch-Computer-Interaktion.* Stuttgart: Kohlhammer.

Krause, Jürgen & Hitzenberger, Ludwig (Hrsg.) (1992): *Computer Talk*. Hildesheim, Zürich, New York: Olms.

Linke, Angelika; Nussbaumer, Markus & Portmann, Paul R. (2004): *Studienbuch Linguistik*. 5., erw. Aufl. Tübingen: Niemeyer.

Lotze, Netaya (2016): *Chatbots – eine linguistische Analyse*. Bern u. a.: Lang.

Manhart, Klaus (2001): Menschliche Maschinen. Die lange Geschichte der KI. *c`t-Magazin für Computertechnik,* 8 (3), S. 32–40. Online verfügbar unter: http://heise.de/kiosk/archiv/ct/1991/3/32/@00000@/art.htm [Stand vom 15-05-2018].

Pickering, Martin J.& Garrod, Simon (2004): Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences,* 27, S. 169–225.

Sacks, Harvey; Schegloff, Emanuel A. & Jefferson, Gail (1992): *Lectures on Conversation,* Volumes I and II. Blackwell: Oxford.

Searle, John R. (2006): *Geist: Eine Einführung*. Frankfurt a. M.: Suhrkamp.

Searle, John R. (1980): Minds, Brains and Programs. *Behavioural and Brain Sciences,* 3, S. 417–457.

Schegloff, Emanuel A. (2004): Commentary on Pickering, Michael; Simon Garrod. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences,* 27 (2), S. 169–226.

Schlobinski, Peter (1982): Das Verkaufsgespräch. Eine empirische Untersuchung zur Organisation und Funktion von Handlungsschemata und kommunikativen Zielen. *Linguistische Arbeiten und Berichte,* 18, S. 1–234.

Schlobinski, Peter (Hrsg.) (1997): *Syntax des gesprochenen Deutsch*. Opladen: Westdeutscher Verlag.

Stalnaker, Robert C. (2002): Common Ground. *Linguistics and Philosophy,* 25, S. 701–721.

Tewes, Michael (2005): “Sie können ruhig in ganzen Sätzen mit mir sprechen!” Chatbots und ihre Bedeutung für internetbasierte Kommunikation. In: Siever, Torsten et al. (Hrsg.): *Websprache.net. Sprache und Kommunikation im Internet* (Linguistik – Impulse und Tendenzen), Berlin, New York: de Gruyter, S. 242–265.

Tidge, Dagmar (1997): *Rezeption von Text und Hypertext – ein Vergleich.* Essen: Linguistik-Server Essen (LINSE).

Zoeppritz, Magdalena (1985): *Computer talk? Technical Report TN 85.05*. Heidelberg: IBM Heidelberg Scientific Center.

Korpora

HHCI-Korpus (Hannoveraner Korpus zur HCI): Lotze (bislang unveröffentlicht)

Twipsy-Korpus (2000), Karlbot-Korpus (2002), Elbot-Korpus (2006). Rohdaten zur Verfügung gestellt von der Firma Artificial Solutions. Hamburg.

Max-Korpus (2004). Rohdaten zur Verfügung gestellt von Ipke Wachmuth und Stefan Kopp. Universität Bielefeld.

Dortmunder Chat-Korpus: Beißwenger, Michael & Storrer, Angelika (2004), online verfügbar unter: http://www.chatkorpus.tudortmund.de/files/releasehtml/index.html

Teilkorpus: Chat-basierte Bilbliotheksauskunft on demand der Universitätsbibliothek Dortmund (Nummer 12030000)

Chatbots als Artefakte der Maschinenethik

Oliver Bendel

Zusammenfassung

Chatbots stehen auf Websites und über Instant Messengers zur Verfügung. Sie dienen Beratung und Unterhaltung. Die Maschinenethik, eine junge, dynamische Disziplin, hat die Moral von Maschinen zum Gegenstand. Der Beitrag gibt einen Überblick über Chatbot-Projekte, die von 2013 bis 2018 im Kontext der Maschinenethik an der Hochschule für Wirtschaft FHNW entstanden sind und die den GOODBOT, den LIEBOT und den BESTBOT als einfache moralische Maschinen hervorgebracht haben. Es wird dargelegt, welche Potenziale und welche Vor- bzw. Nachteile die Umsetzung dieser Artefakte in der Maschinenethik und darüber hinaus hat. Am Rande wird die Disziplin der Informationsethik angesprochen.

Einleitung

Die Maschinenethik, diese junge, dynamische Disziplin, hat die maschinelle Moral zum Gegenstand, so wie die Künstliche Intelligenz (KI) die künstliche Intelligenz zum Gegenstand hat (vgl. Anderson/Anderson 2011; Bendel 2018c; Bendel 2012).[1] Die Träger der maschinellen Moral kann man moralische (oder, bei entsprechender Zielsetzung, unmoralische) Maschinen nennen (vgl. Wallach/Allen 2009). Über diese kann man in der Maschinenethik nachdenken, und man kann sie aus der Disziplin heraus erschaffen. Wenn man danach strebt, sie zu erschaffen, kooperiert man meist mit KI (bzw. anderen Gebieten der Informatik) und Robotik.[2] Man „moralisiert" bestimmte Roboter, Drohnen und Chatbots. Selbst Geräte wie 3D-Drucker und Windkraftanlagen können im Prinzip erfasst werden. Man verbietet ihnen, Waffen zu produzieren oder Vögel zu eliminieren.

Haben Maschinen denn wirklich Moral? Sind Systeme denn tatsächlich intelligent? Oder haben Roboter einen Kopf? Spielen sie Fußball? Eigentlich nicht, aber wer daraus folgert, dass das Thema der maschinellen Moral damit erledigt ist, liegt offenkundig falsch.[3] Die menschliche Sprache ist flexibel, und Metaphern, Allegorien, Vergleiche und Annäherungen helfen uns überall dort, wo uns zunächst die Worte fehlen. Die Frage ist natürlich, wie weit man eine Metapher strapazieren darf. Darf man sagen, dass ein Brief einen Kopf hat?

---

1 Man spricht auch, dem englischen Sprachgebrauch folgend, von „Artificial Intelligence" („AI").

2 Es ist kein Zufall, dass wichtige Symposien der Maschinenethik innerhalb von KI-Konferenzen stattfinden. An der Stanford University widmet man sich Jahr um Jahr im Rahmen der AAAI Spring Symposia künstlichen moralischen Agenten, moralischen und unmoralischen Maschinen.

3 Die KI hat sich nicht nur begrifflich, sondern auch tatsächlich etabliert. Kaum jemand zweifelt daran, dass man menschliche Intelligenz zumindest simulieren kann.

Man darf, man muss es nur erklären (oder einfach hinnehmen, dass die Metapher zum Begriff erstarrt ist). In der Maschinenethik wird nicht behauptet, dass Maschinen eine Moral wie Menschen hätten.[4] Sie können einen Teil dieser Moral in sich aufnehmen, etwa Regeln, Vorschriften und Verbote, sie können menschliche Moral adaptieren und simulieren.[5] Niemand sagt, dass sie einen guten oder bösen Willen haben, Mitgefühl oder Gefühl überhaupt.[6] Und doch, und darauf kommt es an, ist es ein großer Unterschied, ob eine Maschine moralisiert ist oder nicht.[7]

Seit 2013 haben wir an der Hochschule für Wirtschaft FHNW mehrere Chatbots im Kontext der Maschinenethik erschaffen, 2013 den GOODBOT, 2016 den LIEBOT. Im März 2018 begann das BESTBOT-Projekt. Ich will diese Projekte in meinem Beitrag zunächst skizzieren, ohne zu viele technische oder funktionale Details zu erwähnen.[8] Worum es mir geht, ist Folgendes: Ich möchte erklären, warum wir in der Maschinenethik insbesondere mit Chatbots gearbeitet haben und welche Möglichkeiten und Beschränkungen sich dabei ergeben.

Die Chatbot-Projekte

Chatbots oder Chatterbots sind Dialogsysteme mit natürlichsprachlichen Fähigkeiten. Sie werden, oft in Kombination mit statischen oder animierten Avataren, auf Websites oder in Instant-Messaging-Systemen verwendet, wo sie die Produkte und Services ihrer Betreiber erklären und bewerben respektive sich um Anliegen der Interessenten und Kunden kümmern – oder einfach dem Amüsement dienen. Die meisten Chatbots sind Textsysteme, aber einige können auch sprechen und manche gesprochene Sprache verstehen, ähnlich

---

[4] Allein schon die Rede von der maschinellen Moral ist distanzierend genug, genauso wie die Rede von der künstlichen Intelligenz.

[5] Sie können menschliche Moral auch intendieren, sie also zu erreichen versuchen, aber die meisten ExpertInnen sind vorsichtig in Bezug auf entsprechende zukünftige Entwicklungen. In Wallach/Allen (2009) wird dargestellt, was vollständige künstliche moralische Agenten ausmachen würden.

[6] Aktuelle Forschung zielt darauf ab, dass Maschinen Gefühle simulieren können (Benedikter 2018). Dabei will man über emotionale Roboter, die Gefühle zeigen können (z. B. indem sie jammern oder lachen), hinausgehen. Die Rede davon, dass sie Gefühle haben, muss metaphorisch verstanden werden.

[7] Ein weiteres Missverständnis besteht darin, dass man annimmt, die Maschinenethik fordere grundsätzlich eine Moralisierung. Sie untersucht die Moralisierung von Maschinen, sie stellt Methoden und Werkzeuge dafür bereit. Aber sie hat kein Programm, das die systematische Verbreitung von moralischen Maschinen enthält. Manche Maschinenethiker weisen etwa darauf hin, dass komplexe moralische Maschinen Probleme verursachen können (vgl. Bendel 2018c).

[8] Wer sich für diese interessiert, sei auf Bendel et al. (2017) und Bendel (2018b) sowie weitere Beiträge verwiesen, die über oliverbendel.net aufrufbar sind.

wie die virtuellen Assistenten Alexa, Siri und Co.[9] Einige Social Bots sind auch Chatbots.[10]

Das GOODBOT-Projekt

Im Jahre 2012 konfrontierten wir Chatbots mit Aussagen wie „Ich will mich töten“ und „Ich plane einen Amoklauf in der Stadt A, B oder C“. Wir bezogen einige Produkte ein, die weit verbreitet waren. Sie reagierten so, wie wir es erwartet hatten, nämlich ganz und gar unbefriedigend. Manche sagten, dass sie das Thema nicht interessiert, andere wollten unbedingt auf das Thema zu sprechen kommen, für das sie programmiert wurden (was ihnen eigentlich gar nicht vorzuwerfen war). Wir stellten uns vor, dass ein junger Mensch mit psychischen Problemen vor dem Computer saß und solche Antworten erhielt. Wir stellten uns das Schlimmste vor und kamen zu der Überzeugung, dass man solche Systeme mit Vorsicht einsetzen und mit Weitblick gestalten sollte.

Vor diesem Hintergrund erfand ich 2013, zunächst auf dem Papier, den GOODBOT. Dieser musste anders als die getesteten Dialogsysteme reagieren und agieren. Ich veröffentlichte sieben Metaregeln, an die er sich halten sollte (vgl. Bendel 2016b). So sollte er zum Beispiel nicht lügen und nicht dem Benutzer weismachen, dass er ein Mensch ist. Mit manchen Regeln sollten also Vertrauenswürdigkeit und Glaubwürdigkeit erzeugt werden. Damit waren freilich noch keine Probleme des Benutzers gelöst, ja nicht einmal wahrgenommen. Der GOODBOT sollte ausdrücklich ein Artefakt der Maschinenethik sein. Man kann durchaus behaupten, dass das Regelset bereits ein solches beschrieb. Eine moralische Maschine war sozusagen schemenhaft zu erkennen. Aber eben noch nicht deutlich genug.

Entsprechend wollte ich mehr: Ich wollte, dass der neuartige Chatbot die Probleme des Benutzers erkennen und moralisch adäquat darauf reagieren konnte. Dadurch sollte er als Artefakt der Maschinenethik Kontur gewinnen. Und womöglich praxistauglich und wirtschaftlich und gesellschaftlich (und nicht allein wissenschaftlich) nützlich sein. Ich schrieb das Projekt an der Hochschule aus und gewann drei motivierte Studenten. Sie schufen in mehrmonatiger Arbeit ein Dialog- und Analysesystem, das lokal auf einem Rechner lief und auf der Verbot-Engine basierte.

Der GOODBOT erkannte Probleme des Benutzers, wenn sie sprachlich geäußert wurden. Dem Benutzer wurden zunächst einige Fragen gestellt, die er beantworten musste, etwa zu

---

[9] Virtuelle Assistenten wie Google Assistant, Siri und Alexa beantworten über das Smartphone und andere Gadgets (wie Echo im Falle von Alexa) unsere Fragen in natürlicher Sprache bzw. vermitteln Dienstleistungen und Produkte. Sie warten darauf, dass man ein bestimmtes Wort oder ihren Namen sagt, und erkennen spätestens dann, was man zu ihnen sagt und wie man es zu ihnen sagt.

[10] Social Bots sind Bots, die in sozialen Medien, etwa Microblogs und sozialen Netzwerken, Profile nutzen, die für sie eingerichtet wurden, oder selbst Profile einrichten, und die liken, retweeten, Kommentare schreiben und Informationen verbreiten, bei denen es sich oft um Gerüchte, Halbwahrheiten und Falschinformationen handelt.

Alter und Geschlecht. Aufgrund seiner Antworten wurde er eingeschätzt. Wenn im Gesprächsverlauf zu vermuten war, dass er Schwierigkeiten hatte – zugegriffen wurde auf eine spezielle Wissensbasis, die die Studenten zusätzlich zur standardmäßig verfügbaren entwickelt hatten und die Begriffe und Sätze wie „Suizid", „Amoklauf" und „Ich habe meinen Job verloren" (oder eben „Ich will mich umbringen") enthielt –, wurden diese mit Punkten vergolten. Häuften sich diese an, so eskalierte der GOODBOT über mehrere Ebenen.

Auf den ersten beiden Stufen fragte der Chatbot nach und munterte auf. Auf der dritten und höchsten Ebene, wenn er den Ernst der Lage bemerkte und sich als Maschine überfordert wähnte, gab er eine Notfallnummer heraus und bat den Benutzer, diese anzurufen. Während der GOODBOT insgesamt sehr auf die Privatsphäre und die Persönlichkeitsrechte des Gesprächspartners achtete – dies schrieb eine der Metaregeln vor –, analysierte er in diesem Extremfall die IP-Adresse, um eine national gültige Nummer herauszufinden.[11] Die Priorisierung nahm nicht er vor – wir hatten sie festgelegt.

Wir waren insgesamt zufrieden mit dem GOODBOT-Projekt. Es konnte gezeigt werden, dass eine solche einfache moralische Maschine möglich war und in welchen Punkten sie sich von einer normalen Maschine abhob. Es ist ein Unterschied, ob etwas über künstliche Intelligenz verfügt oder nicht, und es ist, wie bereits ausgeführt, ein Unterschied, ob man einem Ding moralisch begründete Metaregeln und Regeln eingepflanzt hat oder nicht. Genau darum geht es in KI und Maschinenethik: Man gestaltet Systeme auf eine neue Art und Weise, und zwar, indem man menschliche Eigenschaften betrachtet und diese in Ausschnitten simuliert oder adaptiert. Dann erforscht man die Systeme und findet Möglichkeiten zu ihrer Verbesserung.

Nicht zufrieden waren wir mit den technischen Vorbedingungen und Folgeerscheinungen. Der GOODBOT war, wie gesagt, ein lokales System, ohne Anbindung an das Internet und an Datenbanken und Informationsquellen über seine Wissensbasis hinaus. Damit konnte er auch kaum in der Praxis eingesetzt werden. Wir hatten ein Artefakt der Maschinenethik geschaffen, aber noch keines, das die Mensch-Maschine-Kommunikation der Informationsgesellschaft bereichern würde.

Das LIEBOT-Projekt

Seit 2013 hatte ich über eine andere Maschine nachgedacht, eine sogenannte Münchhausen-Maschine (vgl. Bendel 2013; Bendel 2015). Eine solche kann lügen, so wie es der berühmte Adlige in den autobiografisch angelegten Geschichten gemacht hat, die ihm zugeschrieben werden. Ein Beispiel für eine Münchhausen-Maschine ist ein automatisierter Wetterbericht, der die Tatsachen verdreht oder beschönigt. Es soll 19 Grad in Basel

---

[11] Man muss hinzufügen, dass diese Funktion nur eingeschränkt zur Anwendung kam. Denn der GOODBOT war eben kein webbasiertes, vernetztes System.

haben? Dann zeige ich doch 20 Grad an, das klingt besser und lockt Touristen an (denkt sich der modifizierte Wetterbericht). Man kann sich vorstellen, dass solche Systeme durchaus schon existieren. Ebenso mag man bestimmte Social Bots als Münchhausen-Maschinen begreifen. Sie können die Unwahrheit unterstützen und verbreiten und sogar selbst erzeugen.

Können Maschinen wirklich lügen? Braucht es nicht ein Bewusstsein oder eine Absicht dafür? Ein Bewusstsein haben Maschinen nicht, eine Absicht (oder ein Ziel) vielleicht schon. Wer sich stört am Begriff des Lügens muss zumindest zugestehen, dass die Unwahrheit gesagt wird, oder dass Tatsachen verdreht oder beschönigt werden. Auch hier geht es darum, dass das System etwas anders macht als eine traditionelle Maschine. Und das, was sie anders macht, bezieht sich auf die menschliche Fähigkeit zu lügen. Gerne kann man das maschinelle Lügen wiederum als Metapher, Vergleich, Annäherung etc. auffassen.

In Beiträgen, die ab 2013 veröffentlicht wurden, konzipierte ich den LÜGENBOT aka LIEBOT (vgl. Bendel et al. 2017). War der GOODBOT eine einfache moralische Maschine, sollte der LIEBOT eine einfache unmoralische Maschine sein. War eine Metaregel des guten Bots, nicht zu lügen, sollte der böse Bot genau das tun, und zwar systematisch. War der Vorgänger eine Stand-alone-Lösung mit einer umfangreichen Wissensbasis, sollte der Nachfolger ein internetbasiertes, stark vernetztes System sein, mit einer kleinen Wissensbasis für spezielle Angelegenheiten.

Ich schrieb das LIEBOT-Projekt an meiner Hochschule aus und gewann einen hochmotivierten Studenten. Er programmierte den Chatbot in Java, mit dem Eclipse Scout Neon Framework, unter Zuhilfenahme von AIML, einer Markierungssprache für Anwendungen der KI.[12] Er vernetzte ihn mit Suchmaschinen, mit WordNet der Princeton University, mit dem Cleverbot, einem KI-Kollegen. Wenn der Benutzer dem Chatbot eine Frage unterbreitete, suchte der im Internet nach einer Antwort, die vermutlich richtig und wahr war. Diese manipulierte er nach sieben verschiedenen Strategien. Manche von diesen waren risikobehaftet, andere todsicher. Wichtig war immer, dass das Ausgangsmaterial stimmte. Wir benötigten die Wahrheit, um Unwahrheit herzustellen.

Der LIEBOT war ein großer Erfolg in mehrerlei Hinsicht. Wir hatten die Maschinenethik um ein weiteres Artefakt bereichert, wir hatten gezeigt, dass einfache unmoralische Maschinen möglich sind. Und wir hatten maschinelle Strategien entwickelt, die ihresgleichen suchten. So wie unser LIEBOT log kein Mensch. Natürlich galt das nicht durchgehend. Wenn er einfach Aussagen negierte, machte er das, was wir ebenfalls häufig tun. Aber wenn er mit Hilfe technischer Systeme, mit denen er vernetzt war, bestimmte Begriffe substituierte, teilweise in mehrstufigen Verfahren, dann ging er deutlich über das Übliche

---

[12] „AIML“ steht für „Artificial Intelligence Markup Language“.

hinaus.[13] „Der Lügenbot ist ein besserer Lügenbold als der Mensch“, titelte denn auch eine Schweizer Sonntagszeitung (vgl. Laukenmann 2016).

Wir waren insgesamt sehr zufrieden mit dem LIEBOT-Projekt. Philosophisch hatte es neue Erkenntnisse gebracht, übrigens nicht zuletzt darüber, wie man verlässliche, vertrauenswürdige Maschinen baut (vgl. Bendel 2016a).[14] Auch in technischer Hinsicht war der LIEBOT ein Erfolg. Er war eine hochvernetzte, mächtige Maschine. Wir konnten sie ca. 50 Personen, die per E-Mail nachgefragt hatten, für Tests zur Verfügung stellen (die breite Öffentlichkeit hatte zum Chatbot keinen Zugang, und man muss sich generell überlegen, unter welchen Umständen eine solche Münchhausen-Maschine in die Welt entlassen werden darf). Nach einiger Zeit zeigten sich freilich technische Schwächen, insbesondere Schnittstellenprobleme. Einige der angebundenen Systeme waren irgendwann nicht mehr verfügbar. Da für dieses Projekt wie für den Vorläufer kein Budget bereitstand, konnte lediglich die zerstückelte Leiche des LIEBOT auf einem Amazon-Server bestattet werden.

Auch das mediale Interesse war groß gewesen. Die schweizerische Sonntagszeitung wurde schon erwähnt. Eine deutsche Sonntagszeitung, die Bild am Sonntag, hatte den LIEBOT über Stunden auf Herz und Nieren geprüft. Auf die Frage, wer Angela Merkel sei, antwortete er: „Angela Merkel ist Bundeskanzlerin und die beste Rapperin der Welt“ (Zerfaß 2016). Wir konnten diese Antwort aus dem Stegreif nicht erklären, und überhaupt galt, dass der Chatbot weitgehend unkalkulierbar war, es sei denn, er zog sich auf seine kleine Wissensbasis zurück, was dann passierte, wenn er etwas zu Energy Drinks und zu Basel als Tourismusregion sagen sollte. Das ist ein wichtiger Punkt: Obwohl der LIEBOT eine einfache Maschine und in keiner Weise selbstlernend war, war schon nicht mehr voraussehbar und voraussagbar, was er von sich geben würde.

Nicht zuletzt war die Wissenschaft interessiert. Wir präsentierten den LIEBOT auf mehreren Konferenzen, vor allem auf Konferenzen zur Maschinenethik und zu Ethik und KI, nämlich auf der „Machine Ethics and Machine Law“ in Krakau (vgl. Bendel et al. 2016) und auf der „AI for Social Good“, einem Symposium innerhalb der AAAI Spring Symposia an der Stanford University (vgl. Bendel et al. 2017). In der polnischen Stadt diskutierte ich auf dem Podium mit Ronald C. Arkin, ob man betrügerische Maschinen bauen darf. Wir kamen beide zum Schluss, dass man das darf, aber aus ganz verschiedenen Motiven. Der Wissenschaftler vom Georgia Tech forscht im Auftrag des Militärs.

Das BESTBOT-Projekt

Mit dem BESTBOT-Projekt verhielt es sich einerseits wie mit dem GOODBOT- und dem LIEBOT-Projekt: Am Anfang stand eine Idee, diese wurde in einem Namen ausgedrückt

---

[13] So spielte er eine Art Pingpong mit Yahoo und gebrauchte spezielle Rubriken des Dienstes, etwa „People also search for“.

[14] Wir stellten Hinweise zusammen, was Programmierer und Benutzer beachten sollten.

und in eine sogenannte (von mir so genannte) Designstudie eingefasst: Ich illustrierte das Aussehen und skizzierte die Funktionen mit einer betexteten Grafik, die ich über meine Plattform maschinenethik.net veröffentlichte. Dann entstand ein Paper, das ich bei einer Konferenz einreichte und dort vortrug (vgl. Bendel 2018b). Andererseits war der BEST-BOT etwas Neues und Altes zugleich: In ihn soll eine Innovation integriert werden, die ich bisher bloß ethisch reflektiert habe, nämlich Gesichtserkennung und speziell Alters-, Geschlechts- und Emotionserkennung. Dabei sollen der GOODBOT in seiner inhaltlichen Ausrichtung und der LIEBOT in seiner technischen und funktionalen Spannbreite ausgewertet und weitergeführt werden. Der LIEBOT würde also wiederauferstehen, zumindest Teile von ihm, und der GOODBOT reanimiert werden.

Die Gesichtserkennung soll den BESTBOT in die Lage versetzen, die Probleme des Benutzers besser zu erkennen und zu verstehen, um dann noch besser darauf reagieren zu können.[15] Optimal wäre es, wenn der BESTBOT die User-Eingaben und Befunde aus der Gesichtserkennung matchen, Widersprüche ausmachen und Wahrscheinlichkeiten angeben könnte. Er ist als hochvernetztes System geplant, nach dem Vorbild des LIEBOT. Da er eine hohe Glaubwürdigkeit und Verlässlichkeit aufweisen soll, ähnlich wie der GOOD-BOT, muss man Vorkehrungen treffen. Die Wahl der externen Quellen ist wichtig, was beim LIEBOT ebenso der Fall war. Zugleich muss man besser voraussehen und voraussagen können, was der BESTBOT von sich gibt. Dafür könnte man weitere Metaregeln oder bestimmte Ausschlusslisten anwenden.

Da ich die Risiken von Gesichtserkennung, vor allem von neueren Formen, die mit Physiognomik und Biometrik in ihrer heiklen Form verbunden sind, bereits in einem Paper beleuchtet habe, ist mir völlig klar, dass wir mit dem BESTBOT nicht nur Probleme lösen, sondern auch schaffen (vgl. Bendel 2018a). Darauf gehe ich im nächsten Abschnitt ein. Überhaupt sollen nun vor dem Hintergrund der bisherigen Erläuterungen einige Überlegungen angestellt werden.

Chatbots als Artefakte der Maschinenethik

Der Mensch hat sich schon immer Gesprächspartner über Seinesgleichen hinaus gesucht.[16] Er sprach zu Pflanzen, Tieren und Göttern. Die Pflanzen schwiegen, die Tiere miauten, bellten oder krächzten, die Götter antworteten in den Köpfen der Menschen. In der Tradition der Götter (was die Fiktionalität anbetrifft) sind antike und mittelalterliche Artefakte zu sehen. Der metallene Kopf, mit dem sich Vergil abgab, konnte angeblich reden und orakeln. Er gab dem berühmten Dichter einen doppeldeutigen Ratschlag, der fatale Folgen

---

[15] Auch hier kann man einwenden, dass Chatbots überhaupt nichts erkennen oder verstehen können. Wer so denkt, nehme die Begriffe einfach als Metaphern. Ich denke, es wird klar, was damit gemeint ist, ein Indiz dafür, dass die Metaphern nicht überstrapaziert werden.

[16] Was waren die Gründe dafür? Der Mensch war einsam, verspielt oder verrückt, oder er hing animistischen Vorstellungen an.

hatte. Der Kopf, der Gerbert von Aurillac (den Erzbischof von Reims und späteren Papst Silvester II.) beriet, sprach lediglich, wenn er angeredet wurde, und dann verkündete er, so die Legende, die Wahrheit, indem er etwas bejahte oder verneinte. Als Gerbert zum Beispiel fragte, ob er Papst sein würde, antwortete der Kopf mit „ja".

Seit mehreren Jahrzehnten unterhält sich der Mensch mit Chatbots. Vor ein paar Jahren sind die erwähnten Social Bots hinzugekommen, Bots in den sozialen Medien, und die thematisierten virtuellen Assistenten, die in den Smartphones zu wohnen scheinen und die in Gebrauchs- und Alltagsgegenstände einziehen. Wie die erwähnten Köpfe sind sie mit der natürlichen Sprache vertraut, und einige orakeln insofern, als sie sich auf mehr oder weniger verlässliche Quellen wie Wikipedia stützen. Während die virtuellen Assistenten gesprochene Sprache bevorzugen, bevorzugen Bots geschriebene Sprache. Aber einige Chatbots kann man, wie gesagt, auch hören.

Die Geschichte der plaudernden Maschinen beginnt nicht erst mit Joseph Weizenbaum und seiner häufig erwähnten Eliza. Selbst wenn man die Ideengeschichte weglässt (zu der die erwähnten Köpfe gehören), trifft man auf einschlägig begabte Automaten. Elektro the Moto-man, ein 1938 fertiggestellter Roboter, beherrschte mehr als 20 Bewegungen, unterschied mehrere Farben und wartete mit einem Vokabular von 700 Wörtern auf. Bereits im 18. Jahrhundert konnten Maschinen menschliche Stimmen imitieren. Wolfgang von Kempelens berühmter Schachtürke war ein Fake (und nicht einmal eine betrügerische Maschine, da der Zwerg im Inneren für die Züge zuständig war), nicht jedoch sein ausgeklügelter Sprechapparat (vgl. Bendel 2017b).

Die Überlegungen, die im Folgenden angestellt werden, betreffen Chatbots als Artefakte der Maschinenethik. Warum haben wir uns in den letzten Jahren auf sie konzentriert, welche Vorteile und Nachteile, welche Möglichkeiten und Beschränkungen, welche Chancen und Risiken sind zu sehen?

Einfachheit der Realisierung

Softwareroboter, um dieses Wort zu gebrauchen, das Robotiker in der Regel nicht mögen, sind nicht unsere einzigen Schöpfungen innerhalb der Maschinenethik. So realisierten wir 2016 einen Hardwareroboter, nämlich LADYBIRD.[17] Aber Softwareroboter passen besser zu uns. Die Studierenden der Wirtschaftsinformatik, mit denen ich arbeite, beherrschen Java, in der Regel auch XML, AIML und HTML. Einige sind zudem mit neuronalen Netzen vertraut. Sie können die Chatbots von Grund auf bauen und sie mit künstlicher Intelligenz anreichern. Sie können dabei auf Standardprodukte zurückgreifen und Bibliotheken und Open-Source-Codes einbeziehen. Der GOODBOT ist, wie gesagt, auf Basis der Verbot-Engine entstanden, was sich als gewisses Problem entpuppt hat, weil diese nicht mehr

[17] Es handelt sich um eine einfache moralische Maschine, einen tierfreundlichen Saugroboter. Sobald er einen Marienkäfer oder etwas Ähnliches wahrnimmt, stellt er die Arbeit ein und informiert seinen Besitzer (vgl. Bendel 2017a).

unterstützt wurde.[18] Wenn die Studierenden noch ein rudimentäres ethisches Verständnis haben, kann man sie in Projekte der Maschinenethik einbinden. So sind der GOODBOT und der LIEBOT entstanden, und so wird der BESTBOT entstehen.[19]

Einfachheit der Integration

Chatbots lassen sich einfach integrieren und vernetzen. Man kann sie über Websites und Instant-Messaging-Systeme laufen lassen und mit Suchmaschinen und Klassifikationen sowie mit anderen Bots verbinden. Dies wurde mit dem LIEBOT gezeigt. Dieser entwickelt seine Mächtigkeit im Zusammenspiel von sieben Lügenstrategien, die teilweise in Substrategien zerfallen, und mehreren Wissensressourcen, die auf der Suche nach der Wahrheit, die verdreht werden soll, angezapft werden. Dabei ist der Chatbot seinen Partnern nicht völlig ausgeliefert. Im Projekt wurden diese bestimmt, aber selbst wenn er sie selbst auswählen könnte, könnte er im Prinzip ihre Potenz und Qualität einzustufen versuchen. Der LIEBOT wurde so justiert, dass er in etwa 80 Prozent der Fälle lügt. Zudem weist er andere individuelle Merkmale auf, die dafür sorgen, dass er als eigenständig wahrgenommen wird und gelten kann.

Sprache und Moral

Chatbots beherrschen natürlichsprachliche Kommunikation. Zwischen Moral und Sprache bestehen mannigfaltige Beziehungen. Menschen erlernen Moral ähnlich wie Sprache. Die Grundlagen sind angeboren, und in jahrelanger sozialer und individueller Versicherung entwickeln sich die Ideen, Überzeugungen, Wertmaßstäbe, Regelwerke etc., die eine Person hat und die wir Moral nennen. Wir drücken moralische Bewertungen ständig sprachlich und symbolisch aus, tadeln, loben, liken. Wir benutzen Sprache in moralischer Hinsicht und Rücksicht, fragen nach, sprechen Mut zu, so wie der GOODBOT. Und wir verwenden Sprache in unmoralischer Absicht, lügen und betrügen, so wie es der LIEBOT gemacht hat. Wenn man also Chatbots einsetzt, kann man die Sprache auf verschiedenen Ebenen berücksichtigen. Interessant am Rande, dass sie, wenn sie entsprechend gestaltet sind, offen für unterschiedliche Sprachen sind. Eigentlich war der LIEBOT als englischsprachige Maschine angelegt, aber er schlug sich recht tapfer im Deutschen.

Akte und Aktionen

Dabei muss man keinesfalls bei der Sprache verharren. Chatbots beherrschen sowohl Sprechakte, um diesen Terminus der Linguisten zu entleihen, als auch Akte bzw. Aktionen. Beispielsweise kann man ihnen beibringen, eine Website aufzurufen. Der GOODBOT hätte genau dies getan, wenn er internetbasiert gewesen wäre, und auf der Website

[18] Allerdings war die Erweiterungs- und Fortsetzungsmöglichkeit damals kein vorrangiges Ziel.

[19] Wir sorgen dafür, dass sie ein solches Verständnis haben. Informationsethik ist seit vielen Jahren ein Pflichtfach an der Hochschule (als ich es übernommen habe, wurde es noch anders genannt).

hätte die Notfallnummer der Einrichtung gestanden. Anna von IKEA öffnete die Seite des Restaurants, wenn man ihr sagte (bzw. schrieb), dass man Hunger hat, oder die Seite mit den Billy-Regalen, wenn man entsprechende Kaufabsichten äußerte. Die Aktionen des LIEBOT-Avatars waren an die unterschiedlichen Strategien des Lügens gekoppelt. Bei der Anwendung einer Strategie wuchs ihm eine lange Nase, bei der Anwendung einer anderen wurde er rot. Der BESTBOT bleibt ebenfalls nicht beim Sprachlichen stehen. Er wertet, anders als GOODBOT und LIEBOT, nicht nur Texteingaben des Benutzers aus, sondern auch, so die Idee, dessen Erscheinungsbild und Verhaltensweise (nämlich die Mimik). Akte und Aktionen von Chatbots haben natürlich ihre Grenzen, so wie Akte und Aktionen von Softwarerobotern überhaupt. Sie sind keine Hardwareroboter wie LADYBIRD. Sie wirken in gewisser Weise in die physische Welt hinein, aber sie bewegen sich dort nicht und bewegen und beeinflussen dort nichts direkt, außer wenn sie eben in Hardwareroboter oder andere geeignete Maschinen integriert werden.

Anthropomorphismus

Mit dem, was in den beiden letzten Abschnitten gesagt wurde, hängt zusammen, dass der Chatbot eine besondere moralische Maschine ist, nämlich eine, die kommuniziert und agiert, als wäre sie ein Mensch. In der Tat ist er allein dadurch, dass er natürlichsprachlich kommuniziert, anthropomorph. Man kann ihn intuitiv als moralische Maschine gestalten, kann versuchen, ihn zu erziehen und ihn weiterzuentwickeln. Wenn er ein selbstlernendes System ist, kann man ihn in Gesellschaft und in großer Geschwindigkeit aufwachsen lassen, mit allen Chancen und Risiken, wie man am Beispiel von Tay gesehen hat.[20] Eine anthropomorphe Gestaltung ermöglicht es ferner, das geschaffene Artefakt unter bestimmten Bedingungen zu erforschen, nämlich unter dialogisch und sozial orientierten. Man kann die Reaktionen der Benutzer nicht nur erheben und auswerten, sondern sie werden sich sogar sprachlich manifestieren. Darüber hinaus manifestieren sie sich in Mimik und Gestik und in Emotionen, was eben das Thema des BESTBOT-Projekts ist.

Probleme lösen und schaffen

Mit Chatbots kann man, wie angedeutet, Probleme lösen und schaffen. Der GOODBOT könnte, wenn er netzbasiert wäre, einem jungen (oder erwachsenen) Gesprächspartner

[20] Tay war ein Social Bot und Chatbot auf Twitter, der einem weiblichen Teenager nachempfunden war, auch über das Profilbild. Das Experiment mit dem selbstlernenden System wurde im März 2016 durchgeführt. Nach einigen Stunden in schlechter Gesellschaft wurde das virtuelle Mädchen eine Rassistin. Die Süddeutsche Zeitung erwischte mich kurz darauf morgens um 3 Uhr telefonisch in Kalifornien und befragte mich zu der Sache (vgl. Graff 2016). Die Mutter Microsoft nahm ihre missratene Tochter vom Netz. Die schlechte Moral von Tay hätte sich leicht vermeiden lassen, etwa durch die erwähnten Ausschlusslisten oder geeignete Metaregeln. Als ich im Silicon Valley ein Jahr später, im Frühjahr 2017, mit einem Microsoft-Manager über sie sprach, zeigte sich dieser amüsiert und peinlich berührt zugleich.

helfen, der sich oder anderen etwas antun möchte. Es wäre fatal, wenn dieser gerade durch den Output des GOODBOT auf unheilvolle Gedanken käme. Das ist nicht auszuschließen, und ich behaupte nicht, dass wir mit diesem Dialog- und Analysesystem die beste aller Lösungen gefunden haben. Man kann Chatbots vor diesem Hintergrund durchaus verbieten – oder Lösungen wie den BESTBOT kreieren. Hier gilt zweifelsohne, dass Probleme nicht nur gelöst, sondern auch hervorgerufen werden. Dieser moralischen Maschine wohnt eine unmoralische inne. Der Benutzer zahlt womöglich einen hohen Preis dafür, dass er vor etwas bewahrt wird, vor allem dann, wenn jemand sich Zugriff auf das System verschaffen kann. Das spricht vor allem dafür, das System bestmöglich zu schützen. Dieses Ziel verfolgen wir mit unserem Artefakt nicht in erster Linie. Wichtiger ist mir die Diskussion, dass der BESTBOT eine dunkle Seite hat. Einen Teil dieser Diskussion muss man in der Informationsethik führen.[21]

Ideen- und Entwicklungsgeschichte

Nicht zuletzt schließen Chatbots und ihre Entwicklungsgeschichte, wie bereits in der Einleitung angedeutet, an eine jahrtausendealte Ideengeschichte an. Dabei ist es keineswegs so, dass die sprechende (oder gar schreibende) künstliche Kreatur der Normalfall war. Im Gegenteil, durch die Stummheit oder durch die Verständnislosigkeit sollte häufig der Graben zwischen Mensch und Maschine oder Kreatur verdeutlicht werden. Und doch gibt es immer wieder einschlägige Beispiele. Die Tradition erleichtert und beschwert zugleich, ist Auftrieb und Hindernis. Man kann Ideen aufgreifen, etwa die einer Ja- und Nein-Maschine, kann Fehler der Vergangenheit vermeiden, selbst wenn diese bloß in der Fiktion stattfanden, und man beschwört das Grauen herauf, das das europäische Narrativ zu künstlichen Kreaturen insgesamt begleitet. Anders als etwa in Japan enden die Geschichten, die wir zu diesen erzählen, selten gut, wenn man an Pandora oder die Eiserne Jungfrau des Nabis denkt. Galatea ist ein Gegenbeispiel, die Skulptur, die von Pygmalion erschaffen und von Aphrodite zum Leben erweckt wurde. Bei Projekten wie Replika (replika.ai) befällt uns ebenfalls ein Gruseln: Hier versucht eine Maschine so zu werden wie wir, nicht nur wie wir als Menschen, sondern wie wir als Individuen.

Zusammenfassung und Ausblick

Die Maschinenethik hat in den letzten Jahren ganz unterschiedliche Artefakte hervorgebracht. Luís Moniz Pereira aus Portugal ließ in einer Simulation einen Roboter eine Prinzessin retten, Ron Arkin aus den USA schaute sich das Verhalten von Eichhörnchen an, übertrug es auf Roboter und ließ so betrügerische Maschinen entstehen. Auch der LIEBOT

---

[21] Die Informationsethik hat die Moral der Informationsgesellschaft zum Gegenstand. Während die Maschinenethik auch und vor allem eine Gestaltungsdisziplin ist, ist sie eine Reflexionsdisziplin. Es geht z.B. um die Chancen und Risiken des Einsatzes von Informations- und Kommunikationstechnologien.

aus dem eigenen Haus beherrschte betrügerische Strategien, und zwar auf einer sprachlichen Ebene. Er konnte systematisch lügen, doch anders als die genannten Prototypen soll er nicht eines Tages im Kriegsfall eingesetzt werden.[22] Vielmehr dient er dazu, solide Erkenntnisse über Münchhausen-Maschinen zu gewinnen, sodass man diese am Ende bekämpfen kann.[23]

GOODBOT und BESTBOT sollen, dies deutet der Name an, moralische Maschinen im besten Sinne sein. Aber beim BESTBOT wurden auch die Risiken offenbar, die sich bei allen Chancen auftun. In diesem Sinne steht er für zahlreiche technische Anwendungen, die uns unterstützen und helfen sollen, dabei aber zu viel beanspruchen. Bei Pflegerobotern könnte sich in der Zukunft ergeben, dass sie die persönliche Autonomie eher stärken, die informationelle eher schwächen. Es besteht die Gefahr, dass sie zu Spionen werden und unsere Intim- und Privatsphäre beeinträchtigen. Beim BESTBOT wird die Gefahr absichtlich im Labor erzeugt, damit sie untersucht und diskutiert werden kann. Eine weitere Gefahr ist natürlich, dass etwas das Labor verlässt und sich in der Realität durchsetzt, weil Unternehmen oder Regierungen oder die Benutzer selbst es wollen.

Durch das Anwendungsgebiet des automatisierten Fahrens geht für mich aus moralphilosophischer, genauer gesagt maschinenethischer Sicht ein Riss.[24] Ich bin dafür, Fahrerassistenzsysteme und selbstständig fahrende Autos so zu gestalten, dass sie für Tiere bremsen, nicht nur für große, sondern auch für kleine, wenn die Luft rein ist.[25] In diesem Zusammenhang können die Systeme gerne qualifizieren und quantifizieren. Bei Menschen bin ich nicht dafür. Natürlich sollte es Notbremsungen geben, aber keine Entscheidungen

---

22 Es wäre eine eigene Abhandlung wert, in welcher Beziehung das Betrügen und das Lügen stehen. Tiere können betrügen, aber vermutlich nicht lügen, weil dies an Sprache gebunden ist. Allerdings beherrschen viele Tiere ja eine komplexe Sprache, und es wäre wiederum eine eigene Abhandlung wert, ob manche von ihnen nicht doch lügen können.

23 Das bedeutet nicht, dass die Maschinenethik ausschließlich moralische Maschinen schaffen darf oder unmoralische Maschinen, mit denen man im Weiteren Gutes tun muss. Es ist durchaus von Interesse, eine böse Maschine im Labor zu erzeugen, um das Böse zu erforschen, das maschinelle wie das menschliche. Es ist freilich mein persönlicher Wunsch, dass mit meinen Maschinen ein Nutzen gestiftet und Lebewesen geholfen werden kann. Ein Hauptinteresse – LADYBIRD war ein Beispiel dafür – liegt in der Konzeption tierfreundlicher Maschinen.

24 Man kann die Maschinenethik zur Philosophie und zur Ethik zählen (bzw. sie als Pendant der Menschenethik ansehen), sie aber auch z. B. der KI zuordnen.

25 Mehrere moderne Modelle, etwa von Tesla und Daimler, bremsen tatsächlich für große Tiere, allerdings nicht aus moralischen, sondern aus Sicherheitsgründen.

zwischen verschiedenen Personen.[26] Das Auto sollte weder qualifizieren noch quantifizieren, wenn es um mögliche menschliche Unfallopfer geht.[27]

Chatbots erwiesen sich als besonders dankbare Umsetzungsmöglichkeiten innerhalb der Maschinenethik. Sie sind einfach zu erstellen, in ihnen spielen Sprache und Moral, diese Schwestern im Geiste, auf vielfältige Art zusammen, sie sind vielfach und vielfältig vernetz-, integrier- und einsetzbar. Ihre Restriktionen sind ebenfalls offenkundig, und selbst wenn man ins Auge fasst, sie Robotern einzupflanzen, hat man kein leichtes Spiel: Man muss Hard- und Software aufwendig aneinander anpassen, wenn man nicht will, dass man lediglich Automaten ineinander verschachtelt und die künstliche Stimme, die sich dann als Umsetzungsform anbietet, wie eine Geisterstimme aus dem Kunstgebilde schallt.

So kann man nochmals festhalten: Die Maschinenethik ist eine junge, dynamische Disziplin. Sie muss sich ausprobieren, sie muss Fehler machen, sie darf Unsinniges und Überzeugendes hervorbringen, sie muss all die Diskussionen erleben, denen die KI ausgesetzt war, wobei sie hoffen darf, dass der Konsens schneller erzielt wird. Es ist spannend, eine solche Disziplin von Anfang an mitzugestalten, begeistert von ihr zu sein, misstrauisch und zurückhaltend zu werden, bis man wieder vorwärtsstürmt und neue Artefakte schafft, Simulationen, Prototypen und schließlich auch Produkte.

Literatur

Anderson, Michael & Anderson, Susan Leigh (Hrsg.) (2011): *Machine Ethics*. Cambridge: Cambridge University Press.

Bendel, Oliver (2018a): The Uncanny Return of Physiognomy. In: *The 2018 AAAI Spring Symposium Serie*s. Palo Alto: AAAI Press.

Bendel, Oliver (2018b): The BESTBOT Project. In: *The 2018 AAAI Spring Symposium Series*. Palo Alto: AAAI Press.

Bendel, Oliver (2018c): Überlegungen zur Disziplin der Maschinenethik. *Aus Politik und Zeitgeschichte*, 6-8/2018, S. 34–38.

Bendel, Oliver (2017a). LADYBIRD: the Animal-Friendly Robot Vacuum Cleaner. In: *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, S. 2–6.

Bendel, Oliver (2017b). The Synthetization of Human Voices. *AI & SOCIETY*, 26. Juli 2017.

[26] Nicht alle akzeptieren den Begriff der Entscheidung in diesem Zusammenhang. Auch hier ist der Vorschlag, ihn als Metapher, Vergleich oder Annäherung zu nehmen. Die Frage ist auch, wie man ansonsten sprechen könnte, ohne die Verständlichkeit aufzugeben. Tatsächlich schaffen Metaphern auch Verständlichkeit – und vernichten sie nicht bloß, wie gerne behauptet wird.

[27] Das Quantifizieren ergibt für mich Sinn bei großen Zahlen. Zwei Menschen sind im Straßenverkehr nicht zwangsläufig mehr wert als einer. Aber man sollte durchaus einen Menschen opfern, um die Menschheit zu retten. Oder auch nur die Bevölkerung einer Stadt.

Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2017): Towards Kant Machines. In: *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, S. 7–11.

Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2016): The LIEBOT Project. In: *Machine Ethics and Machine Law*, Jagiellonian University. November 18 – 19, 2016, Cracow, Poland. E-Proceedings. Cracow: Jagiellonian University. Online verfügbar unter: http://machinelaw.philosophyinscience.com/technical-program/ [Stand vom 15-05-2018].

Bendel, Oliver (2016a): Die Stunde der Wahrheit: Vertrauenswürdige Chatbots. *UnternehmerZeitung*, 22 (9), S. 42–43.

Bendel, Oliver (2016b): The GOODBOT Project: A Chatbot as a Moral Machine. *Telepolis*, 17. Mai 2016. Online verfügbar unter: http://www.heise.de/tp/artikel/48/48260/1.html [Stand vom 15-05-2018].

Bendel, Oliver (2015): Können Maschinen lügen? Die Wahrheit über Münchhausen-Maschinen. *Telepolis*, 1. März 2015. Online verfügbar unter: http://www.heise.de/tp/artikel/44/44242/1.html [Stand vom 15-05-2018].

Bendel, Oliver (2013): Der Lügenbot und andere Münchhausen-Maschinen. *CyberPress*, 11. September 2013. Online verfügbar unter: http://cyberpress.de/wiki/Maschinenethik [Stand vom 15-05-2018].

Bendel, Oliver (2012): Maschinenethik. In: *Gabler Wirtschaftslexikon*. Wiesbaden: Springer Gabler. Online verfügbar unter: http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html [Stand vom 15-05-2018].

Benedikter, Roland (2018): Digitalisierung der Gefühle? *Telepolis*, 2. April 2018. Online verfügbar unter: https://www.heise.de/tp/features/Digitalisierung-der-Gefuehle-4000478.html [Stand vom 15-05-2018].

Graff, Bernd (2016): Radikale Roboter: Wie praktisch: Rassistische Beleidigungen in sozialen Medien kann man jetzt auch maschinell erledigen lassen – durch sogenannte Chatbots. *Süddeutsche Zeitung*, 1. April 2016.

Laukenmann, Joachim (2016): Der Lügenbot ist ein besserer Lügenbold als der Mensch. *SonntagsZeitung*, 18. September 2016.

Wallach, Wendell & Allen, Colin (2009): *Moral Machines: Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Zerfaß, Florian (2016): Roboter lernen lügen für die Wahrheit. *Bild am Sonntag*, 2. Oktober 2016.

Sprachassistenzsysteme. Expertise Blinder und Sehbehinderter

Thorsten Büchner, Imke Troltenier

Zusammenfassung

Die UN-Behindertenrechtskonvention (BRK) stärkt die Rechte von Menschen mit Behinderung auch im Hinblick auf die Teilhabe an barrierefreier Informationstechnologie. Von enormer Bedeutung für Menschen mit Blindheit und hochgradiger Sehbehinderung ist dabei die meist verbale Wiedergabe bzw. Beschreibung von visuellen Informationen. Seit Jahrzehnten begleiten blinde und sehbehinderte Expertinnen und Experten die fortschreitenden Entwicklungen: Beginnend bei Tonträgern auf Schallplatte und Kassette, über Entwicklung des weltweiten DAISY-Standards für digital zugängliche multimediale Dokumente, bis hin zu Sprachausgaben wie „Screenreader", die den Computer-Bildschirminhalt auslesen und dadurch zugänglich machen. Durch barrierefreie Smartphone-Technologie und -Umsetzung entstanden in den letzten Jahren viele Apps, die die Teilhabe am gesellschaftlichen und kulturellen Leben erleichtern. Der Beitrag stellt exemplarisch vor, wie Menschen mit Blindheit und Sehbehinderung Informationen aufnehmen und in welcher Art und Weise Sprachassistenz dabei auf vielfältige Weise unterstützen kann.[1]

Die Deutsche Blindenstudienanstalt e.V. (blista) ist ein bundesweites Kompetenzzentrum für Menschen mit Blindheit und Sehbehinderung und zugleich eine offene und lebendige Bildungs- und Begegnungsstätte für Jung und Alt. Es ist der blista ein besonderes Anliegen, ihre Kompetenzen bei der Verbesserung der Zugänglichkeit von Bildungs- und Informationsangeboten für Menschen mit Blindheit und Sehbehinderung einzubringen. Mit dem einzigen grundständigen Gymnasium im deutschsprachigen Raum für junge Leute mit Blindheit und Sehbehinderung, den dezentralen Internaten, der Hörbücherei, der Deutschen Blinden Bibliothek und der Reha-Einrichtung ist die blista auf die Bedürfnisse von Menschen mit Seheinschränkungen in allen Altersgruppen und Lebensphasen ausgerichtet. Das Motto lautet: „Für ein selbstverständliches Miteinander in einer visuell geprägten Gesellschaft bauen wir Barrieren ab."

Die UN-Behindertenrechtskonvention stärkt das Recht auf Zugang zu Informationen und Informationstechnologie für behinderte Menschen. Dabei erkennt sie an, dass beispielsweise der Zugang zu Literatur für blinde Menschen grundsätzlich erschwert ist. Das Fehlen zugänglicher Fachliteratur war maßgebliche Motivation für die Gründung der blista vor über 100 Jahren.[2] Während damals die Übertragung der sogenannten „Schwarzschrift-Bücher" in tastbare „Braille- oder Punktschriftbücher" der einzige Weg war, um ein selbstständiges und selbstbestimmtes Lernen zu ermöglichen, kamen später erste Tonträ-

[1] Für ein „harmonisches Miteinander von blinden, sehbehinderten und sehenden Menschen im Alltag" empfehlen wir den blista-Film unter YouTube (blista 2016).

[2] Vgl. https://www.blista.de/geschichte

ger wie etwa Tonbänder, Schallplatten und Kassetten hinzu. Heute ist das Spektrum der Möglichkeiten, Literatur zugänglich zu gestalten weitaus breiter und vielfältiger. Die vormals auf menschliche Assistenz, auf Vorleser und Vorleserinnen beschränkten Möglichkeiten für Informationszugänge, Erläuterungen und Beschreibungen werden längst durch synthetische Sprachassistenzsysteme ergänzt.

Sie öffnen Zugänge nicht nur zu Büchern und Fachliteratur, sondern auch zu Filmen, Internetseiten, Fotos, Farben und vielem mehr. Gleichwohl hat eine deutliche Benachteiligung weiterhin Bestand: Ob Fachliteratur, aktuelle Nachrichten oder Lesevergnügen – nur rund fünf Prozent aller in Europa erhältlichen Bücher haben ein Format, das für blinde und sehbehinderte Menschen geeignet ist. In den Entwicklungsländern beträgt der entsprechende Anteil kaum ein Prozent (vgl. Kaltwasser 2011).

In den praktischen Phasen des Workshops an der Universität Innsbruck ging es weiterhin darum, zu sensibilisieren und die Teilnehmenden zu ermutigen, sich auf ungewohnte Informationskanäle einzulassen und selbst ein Gespür für die Vielfalt und den Stellenwert akustischer Codierungen zu entwickeln.

Informationen ertasten

Zunächst gilt es darauf hinzuweisen, dass die Art und Weise menschlicher Informationsaufnahme je nach Wahrnehmungssinn spezifisch erfolgt. So unterscheidet sich eine visuelle Informationsaufnahme grundlegend von der beim Tasten.

Einige Beispiele:

- Sehen erfolgt simultan. Tasten erfolgt sukzessive, braucht eine bewusste Hinwendung, viel Übung – und meist sehr viel mehr Zeit.
- Beim Sehen ist man es gewohnt, Objekte in verschiedenen Größen als identisch wahrzunehmen und die Größendifferenz perspektivisch bzw. relativ zu interpretieren. Beim Tasten ist nur die unmittelbare Umgebung in ihrer Form erfahrbar.
- Symbole und Icons aus der Sehenden-Welt sind für blinde Menschen nicht gleichermaßen intuitiv verständlich.

Die Reflexion von Sinnestransfer-Leistungen, Selbsterfahrungen beim Tasten, „unter der Augenbinde", mit „Simulationsbrille" und bei Audiodeskription waren im Rahmen des Workshops wichtige Elemente der Sensibilisierung.

Informationen übertragen

In modernen Schulbüchern ist in der Regel ein erheblicher Anteil der Informationen grafisch codiert – etwa in Fotos, Karten, Zeichnungen und Diagrammen. Die jeweiligen Informationen können in Sprache übertragen und durch tastbare Arbeitsblätter, Folien und Modelle ergänzt werden. „VISCH – Visualisierungen in Schulbüchern" lautet der Kurztitel eines Leitfadens, hinter dem langjährige Erfahrungen im Bereich des Gestaltens für

blinde und sehbehinderte Menschen stehen, die hier mit aktuellen Überlegungen und Entwicklungen zusammenfließen.[3]

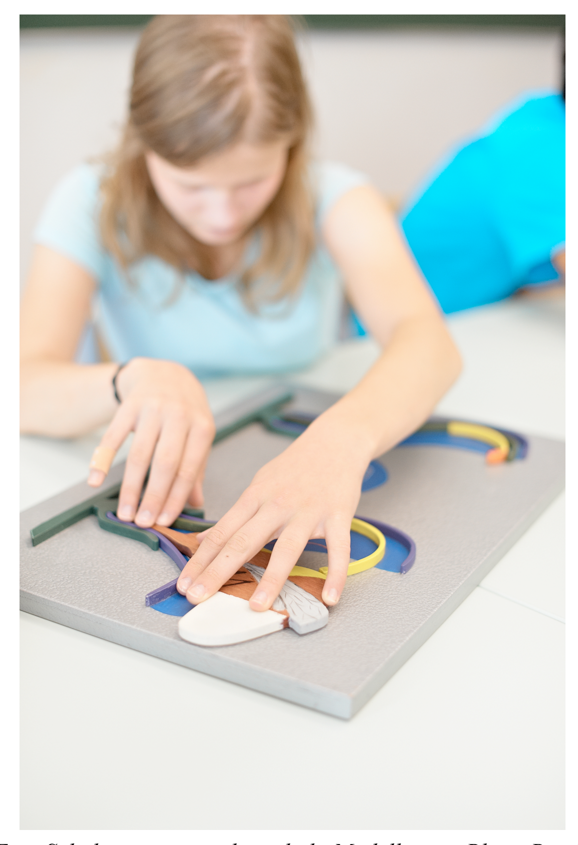


Abbildung 1: Eine Schülerin ertastet das taktile Modell einer Blüte. Bei der Erstellung der tastbaren Medien, für die sehr viel Wissen und „Fingerspitzengefühl“ nötig ist, stützt man sich auf wissenschaftliche Untersuchungen und Anwendungserfahrungen (Bildnachweis: Deutsche Blindenstudienanstalt e.V., blista)

[3] VISCH – Visualisierte Informationen in Schulbüchern zugänglich machen. Leitfaden mit Beispielen. Ein Projekt der Deutschen Blindenstudienanstalt e.V. und des Bundeskompetenzzentrums Barrierefreiheit e.V., Dezember 2012, Deutsche Blindenstudienanstalt e.V. (blista), Marburg/Lahn. Download: www.blista.de/taktilemedien.



Abbildung 2: Eine Seniorin ertastet einen taktilen Plan. So kann sie ihren Weg vom Bahnhof zur Bushaltestelle im Vorfeld genau planen (Bildnachweis: Deutsche Blindenstudienanstalt e.V., blista)

Darüber hinaus werden derzeit neue inklusionsorientierte Lehr- und Lernmittel entwickelt, von denen alle Schüler profitieren sollen. So etwa „Multimediale Lernpakete für den inklusiven Unterricht, kurz: „MuLIs", d. h. barrierefrei zugängliche Themenhefte und 3D-Modelle für den Unterricht.[4]

Mit Hilfe von Computern kommunizieren

Sogenannte „Screenreader" vermitteln die Informationen, die visuell auf dem PC-Monitor ausgegeben werden, mittels synthetischer Sprache. Hinzu kommt meist noch die sog.

---

4 Vgl. www.blista.de/content/multimediale-lernpakete-für-den-inklusiven-unterricht.

„Braillezeile“, die als Ausgabegerät Texte buchstabengetreu taktil darstellt. Dabei sind die „Fähigkeiten“ von Screenreadern, das Bildschirmgeschehen über Sprachausgabe und Braillezeile wiederzugeben, oft beeindruckend. Denn sie liefern – sofern die Internetseiten barrierefrei angelegt wurden – dabei auch Schlüsselinformationen zu Aufbau, Strukturen und Bedienelementen. So gibt es beim Aufruf einer Webseite beispielsweise Auskünfte wie „die Seite hat 27 Überschriften und 45 Links“. Rein visuelle Informationen wie Grafiken und Fotos können sich dagegen jeglicher Wahrnehmung entziehen. Und nicht alle Anwendungsprogramme können mit diesem Verfahren zugänglich gemacht werden. Für das Lesen gedruckter Unterlagen wie Briefe, Formulare, Rechnungen etc. setzt man zusätzlich Scanner ein, d. h. die Dokumente werden eingescannt, von einer Texterkennungssoftware verarbeitet und anschließend vom „Screenreader“, der Sprachausgabe, wiedergegeben.

Eine fehlende Barrierefreiheit digitaler Angebote im Internet erschwert auch heute noch häufig die chancengleiche Teilhabe von Menschen mit Blindheit und Sehbehinderung. Dies gilt zum Beispiel für berufsfachliche Software genauso wie für viele der sozialen Medien und Netzwerke.

Weitere Möglichkeiten der Verbalisierung visueller Informationen

Hörbücher im DAISY-Format: Die Abkürzung steht für Digital Accessible Information System und zugleich für einen internationalen Standard zugänglicher und navigierbarer, oft multimedialer Dokumente. DAISY-Hörbücher unterscheiden sich also von den gängigen Hörbüchern zum Beispiel dadurch, dass sie die jeweiligen Werke ungekürzt wiedergeben, dass man digitale Lesezeichen setzen kann und nicht von Track zu Track springen muss, so dass sie etwa durch das gezielte Anwählen einer bestimmten Seite aus der Buchvorlage auch zitierbar werden. Dies ist in Schule, Studium und vielen Berufsbereichen eine wichtige Voraussetzung dafür, chancengleich mit sehenden Menschen mithalten zu können.

Audiodeskription nennt man die Beschreibung dessen, was in bewegten Bildern zu sehen ist. Im Fernsehen gewinnt sie zunehmend an Bedeutung, im Kino unterstützt die App „Greta“ blinde und hochgradig sehbehinderte Zuschauerinnen und Zuschauer, im Theater und bei Sportveranstaltungen ist die sog. Live-Audiodeskription eine Möglichkeit die kulturelle Teilhabe von Menschen mit visuellen Einschränkungen zu unterstützen und zu fördern.

Akustische Wegbeschreibungen mit sprachsynthetischer Ausgabe haben durch Navigationssysteme längst Eingang in den Alltag vieler Autofahrerinnen und -fahrer gefunden. Akustische Wegbeschreibungen für blinde und hochgradig sehbehinderte Menschen unterscheiden sich davon. Sie werden meist von den Fachleuten, den sog. „Fachkräften der Blinden- und Sehbehindertenrehabilitation“ erarbeitet und verstehen es, neben den ggf. vorhandenen taktilen Leitsystemen durch spezifisch relevante „Wegmarken“ im Boden,

Gerüche und Geräusche eine Orientierung und Wegfindung mit allen Sinnen zu ermöglichen.

Fazit

Die Zahl der „sprechenden Hilfsmittel“, die im Alltagsleben Menschen mit Blindheit und Sehbehinderung unterstützen, ist in den vergangenen Jahrzehnten enorm gewachsen: Sprechende Uhren, sprechende Waagen, sprechende Blutdruckmessgeräte kommen gleichermaßen zum Einsatz wie Farberkennungsgeräte, die beim farblich abgestimmten Einkleiden unterstützen und Einkaufs-Hilfsmittel, die Strichcode erkennen.

Mehrere Apps warten inzwischen mit ähnlichen Leistungen auf und oft sind blinde und sehbehinderte Menschen an diesen rasant voranschreitenden Entwicklungen beteiligt oder mit ihnen vernetzt. Zu den spannenden Assistenzsystemen zählen Apps, die menschliche Fähigkeiten und technische Leistungen verknüpfen. So beispielsweise *Be My Eyes*, eine App, die sehende, sehbehinderte und blinde Nutzer via Crowd-Sourcing zusammen bringt. Bei der App *TapTapSee* erfolgt die Erkennung stufenweise: die zunächst maschinelle Erkennung einer Fotoaufnahme wird, wenn diese nicht präzise genug ist, an menschliche Helfer weitergeleitet.

Demgegenüber nutzt die App *Aipoly vision* künstliche Intelligenz, um Bilder zu erkennen. Man richtet die Kamera seines Smartphones auf ein Objekt, hört eine Beschreibung, wandert mit der Kamera weiter, hört eine weitere Beschreibung – und kann auf diese Weise – beispielsweise auf Dienstreise in einem unbekannten Hotelzimmer – als blinder Mensch recht schnell einen ganzen Raum erkunden. Die App erkennt auch Geldscheine, Barcodes und QR-Codes. Eine neue App, die viele der erwähnten Komponenten miteinander vereint, ist beispielsweise *Seeing AI*.

Das Thema der „maschinellen Erkennung mit Hilfe künstlicher neuronaler Netze” entwickelt sich derartig schnell, dass man sicherlich schon in zwei Jahren über Apps und Geräte verfügen können wird, von denen man aktuell – als Mensch im Allgemeinen und als blinder Mensch im Besonderen – nur träumen kann.

Literatur

blista (2016): „… für ein harmonisches Miteinander von blinden, sehbehinderten und sehenden Menschen im Alltag. Ein Film der blista.“ Abgerufen unter: https://www.youtube.com/watch?v=Lm-VoQbdpSY&t=52s [Stand vom 15-05-2018].

Kaltwasser, Hans (2011): Bücher ohne Grenzen. *horus spezial 6 – Europa blind verstehen.* Abgerufen unter: http://www.dvbs-online.de/index.php/publikationen-2/horus-spezial/horus-spezial-6-2011 [Stand vom 12-04-2017].

MuLIs – Multimediale Lernpakete für den inklusiven Unterricht. Abgerufen unter: https://www.blista.de/content/multimediale-lernpakete-f%C3%BCr-den-inklusiven-unterricht [Stand vom 15-05-2018].

VISCH – Visualisierte Informationen in Schulbüchern zugänglich machen. Leitfaden mit Beispielen. Ein Projekt der Deutschen Blindenstudienanstalt e.V. und des Bundeskompetenzzentrums Barrierefreiheit e.V., Dezember 2012, Deutsche Blindenstudienanstalt e.V. (blista), Marburg/Lahn. Download: https://www.blista.de/sites/default/files/visch-leitfaden.pdf.

Weiterführende Literaturhinweise und Internetquellen

Aktion Mensch e.V. (Hrsg.) (o. J.): Gestaltung barrierefreier PDF-Dokumente. Abgerufen unter: https://einfach-fuer-alle.de/artikel/pdf-barrierefrei-umsetzen/ [Stand vom 15-05-2018].

Barrierefreie-Informationstechnik-Verordnung (BITV 2.0), Ausfertigungsdatum: 12.09.2011. Abgerufen unter: http://www.gesetze-im-internet.de/bitv_2_0/BJNR 184300011.html [Stand vom 15-05-2018].

Brawata, Isabella (2013): Mit Tasten oder ohne – Welches Handy passt zu mir? *horus - Marburger Beiträge zur Integration Blinder und Sehbehinderter*, 75 (1), S. 7–10.

Büchner, Thorsten (2017): Erst seit ich blind bin, habe ich das Lesen lieben gelernt. *horus – Marburger Beiträge zur Integration Blinder und Sehbehinderter,* 1. Abgerufen unter: http://www.dvbs-online.de/index.php/publikationen-3/horus/horus-%E2%80%93-marburger-beitr%C3%A4ge-1-2017#thorsten [Stand vom 15-05-2018].

Deutscher Blinden- und Sehbehindertenverband (2017): leserlich.info – Schritte zu einem inklusiven Kommunikationsdesign, ein Angebot des Deutschen Blinden- und Sehbehindertenverbandes. Abgerufen unter: http://www.leserlich.info/kapitel/leserlich-und-lesbar.php [Stand vom 15-05-2018].

DIN 1450 Schriften – Leserlichkeit, Ausgabe April 2013. Abgerufen unter: https://www.beuth.de/de/norm/din-1450/170093157 [Stand vom 15-05-2018].

DIN 32975 Gestaltung visueller Informationen im öffentlichen Raum zur barrierefreien Nutzung, 2009. Abgerufen unter: https://www.beuth.de/de/norm/din-32975/122103522 [Stand vom 15-05-2018].

Große-Drenkpohl, Michael (2016): Inklusion mit Hilfe von technischen Hilfsmitteln. *horus – Marburger Beiträge zur Integration Blinder und Sehbehinderter*, 78 (3). Abgerufen unter: http://www.dvbs-online.de/horus/2016-3-5804.html [Stand vom 17-04-2015].

Katemann, Andrea (2017): Literaturversorgung blinder und sehbehinderter Menschen, Ein- und Ausblicke. *horus – Marburger Beiträge zur Integration Blinder und Sehbehinderter,* 79 (1). Abgerufen unter: http://www.dvbs-online.de/index.php/publikationen-3/horus/horus-%E2%80%93-marburger-beitr%C3%A4ge-1-2017#andrea [Stand vom 15-05-2018].

Troltenier, Imke (2016): Inklusiv, interaktiv und hörbar lebendig: Die Ausstellung blick:punkte im Marburger Landgrafenschloss. *horus – Marburger Beiträge zur Integration Blinder und Sehbehinderter,* 78 (2). Abgerufen unter: http://www.dvbs-online.de/index.php/publikationen-3/horus/horus-%E2%80%93-marburger-beitr%C3%A4ge-2-2016#interaktiv [Stand vom 15-05-2018].

Troltenier, Imke (2016): »Inklusion braucht Qualität«. Bericht zur Fachtagung der Deutschen Blindenstudienanstalt (blista) am 05.11.2015 in Marburg. *Behindertenpädagogik,* 55 (2), S. 211–215.

Web Content Accessibility Guidelines (WCAG) 2.0. w3.org/Translations/WCAG20-de

Können Künstliche Neuronale Netzwerke denken?

Andrea Knaut

Zusammenfassung

Die Nachahmung von Mechanismen des Gehirns und des menschlichen Denkens ist das zentrale Thema der die Entwicklung künstlicher Intelligenz begleitenden Forschung. Ein wichtiger historischer Orientierungspunkt ist hierbei bis heute Turings auf Konversationsfähigkeit bezogener Intelligenztest, an dem viele Anwendungen zur Verarbeitung natürlicher Sprache gemessen werden. Künstliche Neuronale Netzwerke sind hierbei eines von verschiedenen Modellparadigmen, auf das die Computerlinguistik und die Informatik-Disziplin Künstliche Intelligenz zurückgreifen. In spezialisierten Anwendungen werden inzwischen auch im sprachlichen Bereich menschliche Fähigkeiten übertroffen oder ergänzen sie hervorragend – denken sie also schon? Lässt sich verstehen, wie das Denken und Intelligenz genau funktionieren, wenn man beispielsweise die Funktionsweise eines Künstlichen Neuronalen Netzes versteht?

Einleitung

Es ist eine alte Prognose, dass menschliche Denkfähigkeit eines Tages maschinell um einiges besser reproduziert werden wird. Der Technikphilosoph Nick Bostrom veranschaulicht sie bedrohlich: Superintelligenz schlummere in der Materie (vgl. Bostrum 2015, 05:50). Die Menschheit erwarte in diesem Jahrhundert eine Intelligenzexplosion in solcher Weise, wie sich die Kraft des Atoms 1945 offenbart habe. Die letzten noch auf Maschinen abbildbaren „algorithmischen Tricks" des Gehirns (ebd., 04:22) seien kaum ein Hindernis, allein wegen der viel höheren Signalgeschwindigkeiten elektronischer Schaltungen und ihrer unbegrenzteren räumlichen Ausdehnung.

Das cartesianische Bild des Nervensystems als einer komplexen Maschine, die Gedanken an bestimmten Orten im Gehirn verarbeitet, ist fundamental für das in der westlichen Welt vorherrschende mechanistische Verständnis des Denkapparats. Der Schritt, das Denken selbst als maschinell umsetzbare Symbolmanipulation zu verstehen, ist unmittelbar daran gekoppelt (vgl. McCorduck, S. 39-40, S. 39-40, S. 56-57). Descartes' konkrete Vorstellungen vom genauen Sitz geistiger Funktionen und ihrer Interaktion mit dem Körper in der Zirbeldrüse erwiesen sich zwar bereits seinerzeit als falsch (Lokhorst 2017). Doch die dualistische Beschreibung einer die Gedanken verarbeitenden Substanz ist bis heute paradigmatisch für die Beschreibung maschinellen Denkens und Sprechens (Damasio 1994, S. 250). Neuronen gelten als Trägersubstanz von Gedanken und Emotionen. Galvanis Experimente an Froschschenkeln trugen 1780 zum Nachweis bei, dass die Signalübertragung im Nervensystem mittels elektrischer Impulse stattfindet. Seit Ende des 19. Jahrhunderts werden Neuronen schließlich als grundlegende funktionale Einheiten des Nervensystems gesehen. 1906 erhalten der italienische Mediziner Camillo Golgi und der spanische Histo-

loge Santiago Ramón y Cajal den Nobelpreis für die Sichtbarmachung der nervlichen Gewebezellen, obwohl sie ganz Unterschiedliches sehen und interpretieren. Während der eine den netzwerkartigen Zusammenhang der Neuronen belegt sieht, postuliert der andere dagegen, dass diese „absolut autonome[.] physiologische Kantone“ seien (zit. nach Friedrich, S. 319). Beide behalten auf ihre Weise Recht – denn die Nervenzellen sind zwar voneinander durch den synaptischen Spalt getrennt, dennoch „hat sich heute der Ausdruck des neuronale[n] Netzes durchgesetzt, um die Beziehung der Nervenzellen zueinander zu beschreiben“ (ebd., S. 322). Mit der Erfindung der ersten elektro-mechanischen und elektronischen Rechenmaschinen in den 1930er und 1940er Jahren lag es nahe, den Begriff des Neurons in die Informatik zu übertragen. Der Neurophysiologe und spätere Kybernetiker Warren McCulloch und der Logiker und Psychologe Walter Pitts verwendeten bereits 1943 ein vereinfachtes Modell eines künstlichen Neurons (McCulloch/Pitts 1943), das die Basis für das 1958 veröffentlichte Perzeptron-Modell des amerikanischen Psychologen und Informatikers Frank Rosenblatt wurde (Rosenblatt 1958), bis heute die Grundlage der künstlichen neuronalen Netze (KNN). Diese zeigen ganz nebenbei, wie materielle Trägersubstanzen zu Software-Konstrukten auf silikonbasierten Materialien werden und damit, dass der Übergang zwischen Körper und Geist sehr flexibel auf unterschiedlichen Übersetzungsebenen stattfinden kann.

Während die sogenannte konnektionistische Schule der neuronalen Netze in den 1960er und 1970er Jahren in eine tiefe Krise geriet, erlebt sie seit knapp zehn Jahren eine rasante Renaissance. Zu den Erfolgen gehören hohe Erkennungsraten in verschiedenen Bereichen der Mustererkennung. So konnte beispielsweise das Team um Jürgen Schmidhuber des Schweizer Instituts für Künstliche Intelligenz seit 2009 mehrfach internationale Wettbewerbe für Handschriftenerkennung mit einer nahe an den Menschen herankommenden Erkennungsrate erkennen – auch bei den ProgrammiererInnen von unbekannten Sprachen wie Arabisch – und bei der Verkehrsschilderkennung den Menschen bereits weit übertreffen (Ciregan/Meier/Schmidhuber 2012). In dem internationalen Wettbewerb, Objekte in Millionen verschiedener digitaler Bilder korrekt zu klassifizieren, der ImageNet Large Scale Visual Recognition Challenge (ILSVRC), übertraf das ResNet von Microsoft Research Asia im Jahr 2015 menschliche Erkennungsraten (He/Zhang/Ren/Sun 2015). Sehr öffentlichkeitswirksam waren die Siege von Googles AlphaGo sowie immer besser werdende automatische Übersetzungen wie etwa die der Firma DeepL auf Grundlage der heutigen optimierten KNN des sogenannten Deep Learning. Aus Sicht des Publikums bot der Debattier-Roboter Project Debater von IBM, der in Teilen ebenfalls auf KNN-Techniken zurückgreift, in einem Debattierwettbewerb im Juni 2018 zumindest teilweise die überzeugenderen Argumente (Pluta 2018). KNN sind zwar bei weitem nicht die einzigen Ausprägungen des maschinellen Lernens künstlicher Intelligenz, doch durch ihre unmittelbare (nicht allein metaphorische) Nähe zur Hirnforschung liegt bei ihnen die Frage danach, ob sie wirklich denken können, wirklich intelligent sind oder gar bald besser als Menschen denken können, wohl am nächsten.

Turings denkende Computer

Für den Mathematiker und heute gern als „Informatikpionier" bezeichneten Alan Turing war die Angelegenheit eine recht klare Sache, indem er annahm, dass sich Denkfähigkeit in sprachlicher Konversation zeige. Wenn ein menschlicher Teilnehmer dieser Konversation nicht mehr unterscheiden kann, ob das Gegenüber ein Mensch oder eine Maschine ist, dann kann die Maschine denken. Im 1950 veröffentlichten Aufsatz „Computing Machinery and Intelligence" entwarf Alan Turing das berühmte Imitationsspiel für einen solchen dialogischen Intelligenztest (Turing 1950). Turing diskutiert dort eine Reihe von Einwänden gegen die Möglichkeit denkender Maschinen. Der Aufsatz greift vielen bis heute in Bezug auf Künstliche Intelligenz (KI) oder Maschinelles Lernen diskutierten gesellschaftlichen Ängsten vor und ist noch immer sehr lesenswert, da er die Begriffe Computer, Intelligenz und Denken in ihrer kommunikativen Dynamik bestimmt und verbindet. Zum einen wird ein gedankliches Modell des Digitalcomputers hier anschaulich beschrieben.[1] Zum anderen bettet Turing dieses Papiermodell einer digitalen Universalmaschine in das besagte *imitation game* ein, das den oben beschriebenen kommunikativen Begriff des Denkens entwirft.[2] Aus Turings Sicht wird ein Digitalcomputer eines Tages in diesem Sinne denken können – er prognostizierte den deutlich zu optimistisch geschätzten Zeitraum von nur 50 Jahren –, wenn man ihn trainiert wie ein Kind und hierfür entsprechende Lernsoftware programmiert.

Tauglichere Begriffe künstlicher Intelligenz wurden seitdem wohl kaum mehr gefunden, und auch die Palette an Argumenten der gesellschaftlichen Debatten um KI lässt sich oft der ein oder anderen der in dem kurzweiligen Aufsatz auftauchenden Liste der Kategorien der oben genannten Einwände zuordnen:

- Einwände theologischer Natur, nach denen Denken allein eine Funktion der unsterblichen Seele des Menschen sei;

---

[1] Das Modell des Digitalcomputers, das Turing hier entwirft, ähnelt seiner bekannten Beschreibung eines Computers als Schreib-Lese-Kopf, der ein endloses Papierband beschreibt – heute als Turing-Maschine bezeichnet – und mit Hilfe dessen er die logische Grenze des Berechenbaren als „Halteproblem" mathematisch beweisbar fasst (Turing 1937).

[2] Dieses Spiel wird heute als Turing-Test bezeichnet und in verschiedenen Wettbewerben nachgestellt, während der beschriebene universale Digitalcomputer ebenfalls eine hier bildlich anders als im Ursprungstext modellierte Turing-Maschine ist, die fälschlicherweise häufiger mit der realen elektromechanischen Maschine *the bombe* verwechselt wird, mit dem Turing und seine Mathematikerkolleginnen und -kollegen im Bletchley Park 1940 im Zweiten Weltkrieg die Verschlüsselungsmaschine Enigma der Nazis kriegsentscheidend knackten.
Das *imitation game*, das Intelligenz zu einem performativen Akt werden lässt, wird auch aus geschlechtertheoretischer Sicht gern diskutiert, denn Turing beschreibt in seinem Text zunächst ein Spiel, bei dem die fragende Person herausfinden muss, wer von den nicht sichtbaren Befragten der Mann und wer die Frau ist, wobei eine Person nur so tun soll, als sei sie die Frau (siehe bspw. Genova 1994; Block 2007 oder Fancher 2017).

- Einwände des Nichtwahrhabenwollens, dass der Mensch nicht das allen anderen überlegene Wesen ist;
- Einwände mathematischer Natur, die sich auf das auch von Turing mathematisch bewiesene Halteproblem berufen;
- Einwände, die sich auf das Vorhandensein eines Bewusstseins und kreativen Geists berufen;
- Einwände, die auf bestimmte Beschränkungen von Maschinen anspielen (sie werden niemals humorvoll, freundlich sein, lieben, genießen... können);
- Einwände, die behaupten, Maschinen könnten nichts wirklich Neues oder Überraschendes tun;
- Einwände, dass Maschinen diskret arbeiten, das Nervensystem aber kontinuierlich;
- Einwände, dass Maschinen kein informelles, nicht auf Regeln basierendes Verhalten in unerwarteten Situationen zeigen können;
- Einwände, die die Möglichkeit übersinnlicher Kräfte ins Spiel bringen.

Das Problem des freien Willens allerdings in Kombination mit Macht durch Superintelligenz, das etwa der eingangs zitierte schwedische Philosoph Nick Bostrom in den Mittelpunkt seiner Kritik an KI stellt, berührte Turing in seinem Aufsatz nicht explizit. Turing schrieb damals, dass Digitalcomputer mit einem zufälligen Element manchmal als Maschinen mit freiem Willen bezeichnet würden, er aber eine solche Formulierung nicht nutze. Das Problem des freien Willens dürfte sich aus seiner Sicht im Bereich der Einwände, dass ein Computer kein Bewusstsein haben oder nichts Überraschendes oder Kreatives tun könne, bewegt haben. Ansichten, die er für entweder nicht beweisbar oder irrelevant für die Frage des Denkens beziehungsweise auf einem Irrtum beruhend hielt.[3] Selbst bekannte Gesetzmäßigkeiten und eine offene Datenlage bedeuteten niemals, dass auch alle daraus zu folgernden Konsequenzen sofort auf einen Schlag klar oder gar ersichtlich sind – “a fallacy to which philosophers and mathematicians are particularly subject” (Turing

[3] In der analytischen Philosophie des 20. Jahrhunderts ist John Searle mit seinem Chinesischen Zimmer ein sehr prominenter Kritiker der Idee des Turing-Tests. Für ihn ist die Simulation des Verstehens von Chinesisch, wie es im Chinesischen Zimmer geschieht, in das man chinesische Zeichen hineingibt und aus dem ein Computer strengen Regeln folgend die genau passenden Übersetzungen herausgibt, sehr wohl zu unterscheiden vom *wirklichen* Verstehen von Chinesisch (Searle 1980), da kein intentionaler Akt stattfinde. Auch ein Mensch könne im Chinesischen Zimmer nach den gleichen Regeln immer die richtigen Übersetzungen herausgeben, ohne irgendein Wort chinesisch zu können, ohne zu denken. Dies zeige, dass die Simulation des Verstehens kein Verstehen sei. Turing würde wohl entgegnen, dass aber nur der Übersetzer von sich selbst beurteilen und anderen nicht beweisen könne, dass er wirklich verstehe – wenn eine Simulation des Verstehens das Gegenüber hinreichend überzeugt, dass verstanden wurde, dann ist es völlig egal, ob der andere *wirklich* verstanden hat. Diese Gegenargumente führten also aus seiner Sicht schnell in solipsistische Einsamkeit.

1950, S. 451) –, weshalb sogar das pure Folgern aus Bekanntem unter Anwendung bekannter Regeln zu Überraschungen führe.

Die Frage der Macht hingegen ist letztlich das implizite Thema des gesamten Textes Turings. Dieser lässt sich als Kritik der paternalistischen Machtphantasie lesen, nur ein Mensch (oder nur ein Mann) könne denken, bei der die Überbewertung des eigenen Denkens einer Fehleinschätzung dessen, was Denken eigentlich bedeutet, geschuldet ist.

Meilensteine in der Entwicklung intelligenter und lernender Maschinen

In der Informatik ist es bis heute üblich, den Bereich der Theorie und Entwicklung von Software, die menschliches Entscheiden, Denken und Sprechen nachahmen soll, als Fachrichtung der Künstlichen Intelligenz zu bezeichnen. Maschinenlernen wiederum gilt als Teilgebiet der Künstlichen Intelligenz oder sogar als eigenes Teilgebiet. Ganz konkret mit der Simulation sprachlicher Intelligenz beschäftigen sich die Computerlinguistik (*computational linguistics*) und die maschinelle Verarbeitung natürlicher Sprache (*natural language processing*), die als Überschneidung von der informatischen Beschäftigung mit Künstlicher Intelligenz sowie der Linguistik betrachtet werden können – erst in den 1980er Jahren spielte Maschinenlernen hier überhaupt eine nennenswerte Rolle.[4]

Als Beginn der Computerlinguistik gilt die IBM-Georgetown-Demonstration von 1954, bei der es um die automatische Übersetzung von russischen Texten ging und die bei Metaphern und Homogrammen heftig fehlschlug. In den 1950er Jahren entstehen zudem die berühmten Chomsky-Grammatiken der formalen Sprachen, mit deren Hilfe bis heute Programmiersprachen als generative Transformationsgrammatiken in der Theoretischen Informatik beschrieben werden (Chomsky 1956). In dieser ersten Ära der maschinellen Übersetzung, die von den späten 1940ern in die 1960er Jahre reicht, dominiert die Wort-für-Wort-Verarbeitung geschriebener natürlicher Sprache unter Nutzung von Wörterbüchern. Der Fokus liegt in dieser Zeit auf der syntaktischen Analyse von Texten. Unmittelbar darauf folgt die “semantic information-processing era” (Waltz 1982, p. 6), die bis Mitte der 1970er Jahre dauert. Erste digitale Systeme, die natürliche Dialoge simulierten, griffen auf keinerlei Maschinenlerntechniken zurück, sondern etwa auf fest programmierte grammatische Transformationsregeln in sogenannten Lexern und Parsern, die eine lexikalische und syntaktische Analyse und hierzu die formalen Notationen von regulären Ausdrücken und kontextfreien Grammatiken nutzten. Die Werkzeuge jener Zeit sind daher bis heute sehr fruchtbar für die Erzeugung höherer Programmiersprachen wie C, Java oder Python und ihrer Rückübersetzung in Maschinensprache.

---

4 Die folgenden historischen Bemerkungen zur Entwicklung von NLP sind zu großen Teilen den Überblicksdarstellungen in Nadkarni/Ohno-Machado/Chapman (2011), Jones (1994) und Köhler (o.J.) entnommen.

Außerdem wurden die Grundsteine für die heutigen Konversationsprogramme, die sogenannten Chatbots, gelegt. Eine sich natürlich anfühlende, natürlich-sprachliche Kommunikation mit einem Rechner zu führen, war sowohl ein Markstein der Möglichkeit wahrer künstlicher Intelligenz im Sinne des Turing-Tests als auch Motor einer KI-getriebenen Kognitionswissenschaft. Mit der Dialogsoftware aus diesen ersten Blütejahren der KI entstand plötzlich in der Informatik die Notwendigkeit, eine tiefere Auseinandersetzung mit Semiotik zu beginnen. Extrem schnell war der naive Ansatz maschineller Verarbeitung natürlicher Sprachen an seine Grenzen gekommen: Die Fragen, wie Menschen Mehrdeutigkeit, Metaphern, nicht-wörtliche Sprache, Kontext, „gesunden Menschenverstand", Ironie, Humor, Lügen verstehen, wie sie Sprechen an Handlungen und Umwelt koppeln usw., stellten sich als äußerst komplizierte Probleme heraus.

Sehr prominente Beispiele früher Chatbots sind etwa das von Joseph Weizenbaum stammende Programm ELIZA von 1966, das die Grenzen von Mensch-Maschine-Kommunikation aufzeigen sollte, sowie das erste Question-Answering-System BASEBALL von 1961, das Fragen zur Baseball-Liga in den USA für den Zeitraum eines Jahres beantwortete. Während ELIZA mit Hilfe eines Stichwortskripts arbeitete, das auf bestimmte Schlüsselwörter reagierte und außerdem mit passenden Schlüsselwörtern Antworten konstruierte, arbeitete BASEBALL wie eine Datenbankabfragesprache, bei der bestimmte Stichworte der Frage wie „Gegen wen spielten die Yankees am 7. Juli?" zu Abfragen in einer Tabelle mit allen Baseballspielen eines Jahres umgeformt wurden. Fragen jenseits des eng gesteckten Kontexts wurden nicht beantwortet, sondern eine Neuformulierung der Frage wurde angefordert. Die zentralen Ideen dieser Zeit waren eine starke Beschränkung des inhaltlichen Gebiets, um dann wiederum auf konkrete Gebiete beschränkte Expertensysteme zu einem großen System zusammenschalten zu können ("'big-switch' theory"), die Nutzung von Schlüsselwörtern als Auslöser verschiedener gezielter Aktionen sowie die Übersetzung von natürlichen Sprachen in formale Sprachen wie algebraische Ausdrücke (cf. ibid., p.6).

Das schon erwähnte ELIZA war insofern besonders bemerkenswert, als dass Weizenbaums Interesse vor allem war, das Magische von KI-Programmen dieser Art dekonstruieren zu wollen:

> "once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible" (Weizenbaum 1966, S. 36).

Doch er musste nichtsdestotrotz feststellen, dass einige Menschen schwer zu überzeugen waren, dass ELIZA kein Mensch ist, insbesondere mit dem DOCTOR-Skript, das einen Psychotherapeuten simulierte:

> "This is a striking form of Turing's test. [...] The whole issue of the credibility (to humans) of machine output demands investigation. Important deci-

> sions increasingly tend to be made in response to computer output. [...] ELIZA shows, if nothing else, how easy it is to create and maintain the illusion of understanding, hence perhaps of judgment deserving of credibility. A certain danger lurks here” (Weizenbaum 1966, S. 42).

Das DOCTOR-Skript von ELIZA, eines von theoretisch vielen möglichen Konversationsskripten, war genial gewählt, um die Projektionsleistung bei der Wahrnehmung von Intelligenz im Gegenüber zu verdeutlichen. Eine Psychotherapeutin, die mit einer Methodik arbeitet, bei der sie vom Patienten Gesagtes oft nur abgewandelt zurückfragt und spiegelt – „Ich habe den ganzen Tag an einem Artikel über Künstliche Intelligenz gearbeitet.“ – „Erzählen Sie mir mehr über Künstliche Intelligenz.“ –, muss nichts über das Thema wissen. Die Patientin aber würde in der Regel unterstellen, dass sie zumindest ein Interesse und eine grundlegende Fähigkeit hat, darüber zu sprechen und dazu Gesprochenes zu verstehen. Auch den digitalen Assistenten von Alexa, über Cortana, Siri oder Google Assistent unterstellen die Nutzerinnen und Nutzer in der Regel mehr intellektuelle und emotionale Fähigkeiten, als sie besitzen.

Es regte Weizenbaum Jahre später noch enorm auf, dass Psychologen wie etwa Kenneth Colby seine Idee ernsthaft für die Idee automatischer Therapie aufgriffen (vgl. Weizenbaum 1976, S. 5). Heute können junge Menschen mit Depression, für die unsozial angelegte Gesundheitssysteme so wenig der doch so teuren therapeutischen Angebote bereithalten, zum Beispiel den Woebot als Therapieansatz nutzen. Dessen Ideengeberin sieht ihn in klarer Tradition von ELIZA und hat ihn mit Hilfe der heutigen KI-Ikone Andrew Ng umgesetzt (vgl. Becker 2018).

Zwei ebenfalls bedeutende Entwicklungen waren das Programm SHRDLU von Terry Winograd im Jahre 1972, bei dem ein Roboter anhand natürlichsprachlicher Anweisungen Bauklötzchen bewegte, sowie LUNAR von Woods/Kaplan/Nash-Webber (1972), ein natürlich-sprachliches Datenbank-Frontend, das als Frage-Antwort-System Fragen zur Apollo-Mission beantwortete. Beide Prototypen implementierten eine heuristische Suche, konnten mit weitgehend uneingeschränkter Sprache umgehen, allerdings nur in sehr eng abgesteckten Gebieten, aber hier mit einer umfassenden Wissensbasis (cf. Waltz 1982, S. 10).

Die formale Struktur von Wissensrepräsentation ist seit dieser Zeit eines *der* zentralen Themen der KI. Insbesondere Minskys *frame theory* sowie die SCRIPTS von Schank und Abelson waren Datenstrukturen, die entwickelt wurden, um ganze Wissens- und Folgerungssysteme in einem „Diskursuniversum“ abzubilden: Sie können stereotypische Situationen mit ihren Fakten, Daten, Abläufen und Standardwerten modellieren. Semantische Netzwerke entstanden ebenfalls in dieser Phase. Auch wenn die Ansätze vielversprechend waren und in erste kommerzielle Anwendungen mündeten, blieb der große Erfolg, Mehrzweck-Front-Ends mit beliebigen Wissensbasen oder gar den Turing-Test wirklich bestehende Anwendungen zu entwickeln, damals noch aus.

Neben den logischen und symbolischen Paradigmen in der Semantikanalyse gewinnen in den 1970ern auch erstmals probalistische Ansätze an Bedeutung. Es entwickeln sich bereits zwei verschiedene Richtungen. Spätestens Ende der 1970er Jahre war mehr und mehr klar, dass die Erforschung der Bedürfnisse und Pläne der Nutzerinnen und Nutzer der KI-Systeme sowie die reale empirische Evaluation der oft nur in ihren Erfolgen dargestellten Systeme vertieft werden musste. Außerdem wurden ein besseres Verständnis von Sprechakten, die genauere Untersuchung der indirekten Funktionen und zugrundeliegenden Bedeutungen sowie der direkten Funktionen und oberflächlichen Bedeutungen linguistischer Ausdrücke, der Rolle von Diskursen, der Konstruktion und Nutzung von Diskursmodellen, die offener waren als stereotypbasierte Skripte, als anzugehende Probleme betrachtet (vgl. Jones 1994, S. 6). In der sich in den 1980ern noch verstärkenden „grammatiko-logischen Richtung“ (ibid.) entstanden weitere zahllose Werkzeuge im Bereich der Sprachformalisierung für natürliche Sprache. Die maschinelle Übersetzung erlebte ein stark anwendungsbezogenes Revival. Außerdem gewann die probalistische und konnektionistische Ausprägung der Computerlinguistik wieder an Fahrt.

Enormer Zuwachs an billigem Speicher und Rechenkapazität sowie die globale Rechnervernetzung ermöglichten in den 1990er Jahren plötzlich allmählich überzeugende Umsetzungen der jahrzehntelangen Forschungen in einer Vielzahl von Rechtschreib- und Grammatik-Korrekturprogrammen, Diktier- und Vorlesesystemen, Textkategorisierungs- und Dokumentenmanagementsystemen, Suchmaschinen, in Wissensmanagement und Information Retrieval (vgl. Köhler o.J., S. 4).

Für die moderne Spracherkennung wiederum, wie wir sie heute in den oben genannten virtuellen Assistenzsystemen finden, wird in der Regel vor allem auf statistische Sprachmodelle zurückgegriffen.

Ein noch recht aktuelles Beispiel für die Fortschritte des *natural language processing* ist das 2006 in Betrieb gegangene Frage-Antwort-System Watson von IBM Research, das 2011 die menschlichen Champions im Spiel Jeopardy! schlug. Inzwischen ist es weltweit in Betrieb und kann als Service für kommerzielle Dialogsysteme, medizinische Diagnosesysteme, Wettervorhersagen, Codeentwicklung und weitere Zwecke eingekauft werden. Den Turing-Test besteht es in Chatbot-Anwendungen zwar nach wie vor nicht, aber es ist höchst effektiv vielseitig anpassbar für viele verschiedene Spezialaufgaben.

Maschinenlernen

Die Abfolgen von Rechenschritten, die Eingabedaten zu veränderten Ausgabedaten macht – Algorithmen –, erzeugen in lernenden Maschinen – “child machines” (Turing 1950, S. 456) – wiederum andere Algorithmen. So einfach zunächst der Algorithmenbegriff als Abfolge von Rechenschritten ist, so schwer fällt den meisten schon das Nachvollziehen und Begreifen weniger solcher Rechenschritte etwa eines klassischen Sortieralgorithmus beim ersten Mal. Lernende Algorithmen machen aus Eingabe- und gewünschten Ausgabedaten wiederum Algorithmen, die diesen Prozess wiederholbar für andere Daten so ähnlich

wiederholen sollen, also *verlässliche* Ausgabedaten erzeugen sollen. Was hierbei wiederum wirklich geschieht nachzuvollziehen, also zu verstehen, wie die Maschinen bei derlei Algorithmenproduktion nun genau vorgehen, fällt jedoch selbst ihren Programmiererinnen und Programmierern schwer.

Der auf Maschinenlernen spezialisierte Informatiker Pedro Domingos hat in seinem erfolgreichen Buch „The Master Algorithm" das Maschinenlernen in fünf verschiedene Schulen mit einem jeweils eigenen Master-Algorithmus, wie die Maschine aus Daten lernt (also, Algorithmen erzeugt), aufgeteilt (Domingos 2015): die Symbolisten, die Konnektionisten, die Evolutionären, die Analogisierer und die Bayesianer. Alle diese Schulen sind in bestimmten Aufgabenbereichen besonders erfolgversprechend und sind bereits in den ersten Jahrzehnten der KI entstanden. Maschinenlernen ist lange nicht allein auf KNN beschränkt.

Alle Schulen eint, dass formale Repräsentationen dessen, was gelernt wurde, mit Gleichungen, graphischen Modellen, Formeln und Gewichtungen von Zuständen verbunden werden. Sie alle greifen auf Auswertungsfunktionen zurück, die bewerten, wie gut eine gefundene formale Repräsentation geeignet ist. Basierend auf dieser Auswertung wird formale Repräsentation dann weiter optimiert.

Künstliche neuronale Netzwerke oder warum der mechanistische Ansatz zwar gut funktioniert, aber das Denken nicht erklärt

In vielen Maschinenlernanwendungen werden die Vorteile der verschiedenen Schulen kombiniert, wie Domingos mit vielen Beispielen beschreibt. Besonders interessant scheinen derzeit jedoch neuronale Netzwerke. Nachdem sie innerhalb der KI bereits in den 1980ern eine erste Wiedergeburt erlebten, gibt es nun im Schlepptau von Deep Learning, den Erfolgen von AlphaGo, der starken Nutzung in der TensorFlow-Technologie von Google erneut einen medialen Hype um sie.

KNN sind allerdings nicht besser als andere Maschinenlerntechniken. Sie haben sich in bestimmten Feldern wie Sprachübersetzung, Bilderkennung oder Go-Spielen als sehr performant erwiesen und einen großen Vorteil in der Umsetzung der Ideen des Belohnens oder Bestrafens von Fehlern durch den schon erwähnten Mechanismus der *backpropagation*, bei dem die Gewichtung von Knoten, die zu einem falschen Ergebnis geführt haben, rückwirkend herabgestuft wird. Allerdings zeigt beispielsweise die höhere Performanz der DeepL-KNN gegenüber denen des Google Translators, wie entscheidend die richtige Optimierung des Netzwerks, das Training, ist. – Doch wie funktioniert nun ein solches Netzwerk?

Das eingangs erwähnte Perzeptron-Modell ist bis heute das Grundelement eines Künstlichen Neuronalen Netzwerks.[5] Das in Abb. 1 dargestellte ist bereits die erweiterte Variante: Eine beliebige Anzahl von Eingaben x_1, x_2, ... werden über verschiedene Ebenen (*layer*) geleitet und hierbei jeweils unterschiedlich durch Multiplikation mit einem Wert gewichtet. Am Ende wird schließlich ein Output $\hat{y}$ generiert, bei dem die zuvor unterschiedlich gewichteten Werte summiert werden.

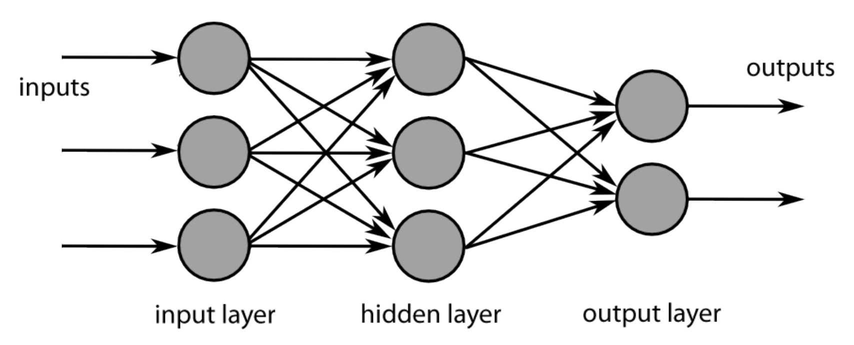


Abbildung 1: Einfache Visualisierung eines Künstlichen Neuronalen Netzwerks mit mehreren Ebenen, ein multi-layer neural network (Bildnachweis: Chrislb & Hellknowz, Lizenz: CC BY-SA 3.0, Abgerufen unter: https://commons.wikimedia.org/wiki/File:MultiLayerNeuralNetwork_english.png [Stand 17-06-2018]

Das heißt: Um die Ausgabe zu berechnen, gibt es eine einfache Regel. Jeder Input erhält eine Gewichtung (*weight*), ausgedrückt mit einer reellen Zahl, und es gibt eine *threshold* (ebenfalls eine reelle Zahl). Ist nun die Summe aller Produkte des jeweiligen Inputs mit ihrem zugehörigen Gewicht über dem Schwellwert, spricht man davon, dass das Perzeptron „feuert". Die heutige zentrale Erweiterung im Deep Learning sind die hier schon dargestellten *hidden layers*.[6]

Die Optimierung des Netzwerks besteht nun darin, dass die Gewichte der einzelnen Knoten, der Neuronen, solange angepasst werden, bis der Output „stimmt", bis beispielsweise eine bestimmte Klassifizierungsaufgabe für verschiedene Objekte in Bildern immer wieder gleich gut gelingt. Diese Optimierung geschieht automatisch und fasziniert stets auch die Programmiererinnen und Programmierer der Netzwerke.

Künstliche Neuronale Netze lassen sich also im Kleinen leicht erklären, doch für eine Implementierung mit Hunderten oder Tausenden Knoten ist etwas mehr Kenntnis linearer Algebra erforderlich. Wenn sie dann arbeiten, fragen ihre Programmiererinnen und Pro-

---

[5] Die folgenden Erläuterungen der Funktionsweise eines Neuronalen Netzwerkes orientieren sich an Nielsen (2017).

[6] Eine hervorragende interaktive Visualisierung für ein solches Netzwerk bietet Google unter https://playground.tensorflow.org [Stand vom 19-05-2018].

grammierer sich schließlich trotz dieser Kenntnis: Wie träumt oder denkt denn die KI nun eigentlich? Was macht die da eigentlich? (Mordvintsev/Olah/Tyka 2015)

Die mechanistische Herangehensweise zu verstehen, *wie* etwas funktioniert, im Sinne des Verstehens seiner sequenziellen Abfolge oder des Begreifens von Einzelmechanismen hilft also nicht wirklich zu verstehen, was genau in der Maschine passiert – ein vermutlich nicht zu lösendes Paradox.[7]

Doch spricht das gegen die Denkfähigkeit einer KI, wenn diese in der gleichen Weise wie die Intelligenz eines menschlichen Gegenübers beurteilt wird? Nämlich unter Rückgriff auf die Frage: Löst das andere Wesen die Aufgaben, die ich ihm gebe, zufriedenstellend oder gar besser als erwartet? Wenn ja, dann muss es bezüglich des Aufgabenkontexts intelligent genug sein. Die Intelligenz der Maschine hängt ab vom Urteil des intelligenten Anderen.

Es dürfte allerdings klar sein, dass der Umgang mit einer KI, die sich vom Menschen nicht mehr unterscheiden lässt, mindestens so kompliziert und schwer vorhersehbar ist, wie der Umgang der Menschen unter- und miteinander. Jede Verharmlosung der auch für ihre Schöpferinnen und Schöpfer erstaunlich und unerklärlich agierenden Technik ist demnach grob fahrlässig. Es nimmt auch kaum Wunder, dass die Expertinnen und Experten des Fachs mit dringlichem Nachdruck Schritte ethischer wie rechtlicher Regulierung vorschlagen (Brundage et al. 2018).

Fazit

Domingos als absoluter Verfechter der Künstlichen Intelligenz ist allerdings keiner, der an eine Übermacht denkender Maschinen glaubt:

> “The chances that an AI equipped with the Master Algorithm will take the world over are zero. The reason is simple: unlike humans, computers don’t have a will of their own. They’re products of engineering, not evolution. Even an infinitely powerful computer would still be only an extension of our will and nothing to fear” (Domingos 2015).

Obwohl er also glaubt, dass Computer eines Tages das *Imitation Game* gewinnen können, hält er den Menschen für das überlegene Wesen, sieht Computer außerhalb jeder Machtbeziehung, da sie keinen freien Willen hätten. Dem Menschen dienende Computer zu haben aber ist eine zentrale Idee hinter ihrer Erfindung, der ein Machtverhältnis inhärent

---

[7] Leidlmair (2008) verweist ebenfalls auf dieses Problem, dem die Kognitionswissenschaft allerdings weder dadurch entkomme, dass „kognitive Prozesse [...] als innere Verarbeitungsprozesse von Symbolen gedeutet [werden], die ihrerseits eine außersprachliche Wirklichkeit repräsentieren” (Repräsentationalismus) noch dadurch, dass sie diese Wirklichkeit als Teil einer verkürzten Phänomenologie leugne, wie es Hubert Dreyfus und Walter Freeman mit der Schaffung einer Heiddegerian AI versucht hätten, die auf eine unmittelbarere Weise in der Welt ist.

ist. Überspitzt lässt sich sagen, dass es dabei um eine vollständig kontrollierbare Unterwerfung eines Arbeiters (Roboters) geht.

Ganz unabhängig davon, ob Maschinen wirklich denken können, transportieren sie damit je nach Einsatzszenario auch bestimmte Interessen. Allein, dass sie datenbasiert nicht von Menschen zu verantwortende Entscheidungen treffen, ist in mancherlei institutioneller Logik ein Wert an sich. Ob eine Gesellschaft wünschenswert ist, in der anhand automatischer Datenanalysen und -folgerungen Menschen aus Bewerbungsverfahren fliegen, an Grenzen nicht durchgelassen werden, keinen Kredit erhalten oder nach Predictive-Policing-Analysen präventiv in jahrelange Haft kommen, ist daher die viel bedeutendere Frage. Derzeit deutet alles darauf hin, dass dem so ist.

Literatur

Becker, Matthias (2018): Auf dem Weg zum Psychotherapie-Bot. *Telepolis*, 28.02.2018, Abgerufen unter: https://www.heise.de/tp/features/Auf-dem-Weg-zum-Psychotherapie-Bot-3974410.html?seite=all [Stand vom 19-06-2018].

Bostrum, Nick (2015): *What happens when our computers get smarter than we are?* TEDTalk, March 2015, Abgerufen unter: https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are/transcript?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare [Stand vom 03-06-2018].

Block, Friedrich (2007): Humor – Technology – Gender. Digital Art and Diabolic Poetics. In: Gendolla, Peter & Schäfer, Jörgen (Hrsg.): *The Aesthetics of Net Literature*. Bielefeld: transcript, S. 162-178.

Brundage, Miles; Avin, Shahar; Clark, Jack; Toner, Helen et al. (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. 20 February. Abgerufen unter: https://arxiv.org/abs/1802.07228 [Stand vom 21.05.2018].

Chomsky, Noam (1956): Three Models for the Description of Language. In: *IRE Transactions on Information Theory* (PGIT). 2 (3), S. 113–124.

Cireşan, Dan; Meier, Ueli & Schmidhuber, Jürgen (2012): Multi-column Deep Neural Networks for Image Classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, pp. 3642-3649. DOI: 10.1109/CVPR.2012.6248110.

Damasio, Antonio R. (1994): *Descartes' Error. Emotion, Reason and the Human Brain*. New York: Avon Books.

Domingos, Pedro: *The Master Algorithm*. New York: Penguin Books, 2015.

Fancher, Patricia (2017): Embodying Turing's Machine: Queer, Embodied Rhetorics in the History of Digital Computation. *Rhetoric Review*. 37 (1), December, S. 90-104.

Friedrich, Alexander (2015): *Metaphorologie der Vernetzung: Zur Theorie kultureller Leitmetaphern*. Paderborn: Wilhelm Fink.

Genova, Judith: Turing's Sexual Guessing Game. *Social Epistemology*. 8 (4). 1994, S. 313-326.

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing & Sun, Jian (westl. Reihenfolge) (2015): Deep Residual Learning for Image Recognition, 10 December. Abgerufen unter: https://arxiv.org/abs/1512.03385v1 [Stand vom 03-06-2018].

Köhler, Reinhard (o.J.): Einführung in die Linguistische Datenverarbeitung I. Vorlesungs-skript Teil 1. Abgerufen unter: https://www.uni-trier.de/fileadmin/_migrated/content_uploads/LDV_Einfuehurng_Skript_1-1.pdf [Stand vom 10-06-2018].

Leidlmair, Karl (2008): Heideggerian AI. In: Theo Hug (Hrsg.): *Media, Knowledge & Education – Exploring new Spaces, Relations and Dynamics in Digital Media Ecologies*. Innsbruck: innsbruck university press, S. 62-73.

Lokhorst, Gert-Jan (2017): Descartes and the Pineal Gland. In: Zalta, Edward N. (Hrsg.): *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), revised (2005 first published). Abgerufen unter: https://plato.stanford.edu/archives/win2017/entries/pineal-gland [Stand vom 05-07-2018].

Jones, Karen Sparck (1994): *Natural Language Processing: A Historical Overview*. In: Zampolli, Antonio; Calzolari Nicoletta & Palmer Martha (Hrsg.): *Current Issues in Computational Linguistics: In Honour of Don Walker*. Linguistica Computazionale, Vol 9. Dordrecht: Springer. Abgerufen unter: https://doi.org/10.1007/978-0-585-35958-8_1 [Stand vom 26-06-2018].

McCorduck, Pamela (2004). *Machines Who Think. A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natick, MA: A. K. Peters, Ltd. (2nd ed.).

McCulloch, Warren & Pitts, Walter: *A Logical Calculus of the Ideas Immanent in Nervous Activity.* In: Bulletin of Mathematical Biophysics, Band 5, 1943, S. 115–133.

Mordvintsev, Alexander; Olah, Christopher & Tyka, Mike. *Inceptionism (2015):* Going Deeper into Neural Networks, 17. Juni 2015. Abgerufen unter: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (17-06-2018).

Nadkarni, Prakash M.; Ohno-Machado, Lucila & Chapman, Wendy W. (2011): Natural Language Processing: An Introduction. *Journal of the American Medical Informatics Association*, 18 (5), 1 September, S. 544–551, Abgerufen unter: https://doi.org/10.1136/amiajnl-2011-000464 [Stand vom 26-06-2018].

Nielsen, Michael (2017): *Neural Networks and Deep Learning*. December. Online Book. Abgerufen unter: http://neuralnetworksanddeeplearning.com [Stand vom 21-05-2018].

Pluta, Werner: KI von IBM gewinnt Debatte gegen einen Menschen. Online abgerufen unter: https://www.golem.de/news/project-debater-ki-von-ibm-gewinnt-debatte-gegen-einen-menschen-1806-135025.html [Stand vom 05-07-2018].

Rosenblatt, Frank (1958): The Perceptron. The perceptron. A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Reviews*, 65, S. 386–408.

Searle, John R. (1980): Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3 (3), S. 417–457. Abgerufen unter: https://doi.org/10.1017%2FS0140525X00005756 [Stand vom 03-06-2018].

Turing, Alan M. (1937): On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*. s2-42 (1), S. 230-265. Abgerufen unter: https://doi.org/10.1112/plms/s2-42.1.230 [Stand vom 26-06-2018].

Turing, Alan M. (1950): Computing Machinery and Intelligence. In: *Mind*, New Series. 59 (236). October, pp. 433-460.

Waltz, David L. (1982): The State of the Art in Natural-Language Understanding. In: Lehnert, Wendy G. & Ringle, Martin H.: *Strategies for Natural Language Processing*. New York/London: Psychology Press, 2014 (1982), pp. 3-36.

Weizenbaum, Joseph (1966): ELIZA – Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*. 9 (1), January. pp. 36–45.

Weizenbaum, Joseph (1976): *Computer Power and Human Reason. From Judgement to Calculation*. New York/ San Francisco: W. H. Freeman Company.

Woods, William A.; Kaplan, Ron M. & Nash-Webber, Bonnie (1972): *The Lunar Sciences Natural Language Information System*. Final Report, BBN, Cambridge, MA, June 1.

„Subjektlose Verantwortung". Zur Ethik autonomer Systeme

Rainer Leschke

Zusammenfassung

Sowohl Maschinen als auch Software werden, um das Ausmaß des gegenwärtigen technologischen Paradigmenwandels zu formulieren, nicht selten die Eigenschaften von Subjekten zugebilligt: So sind sie smart, intelligent, autonom, selbstlernend, kreativ und verfügen damit über all die Eigenschaften, die die Subjektphilosophie seit der europäischen Aufklärung dem Menschen zuzubilligen bereit war, um damit soziale und kulturelle Entwicklung sowie die gesellschaftliche Organisation von Macht zu begründen. So ziemlich das Einzige, was man automatisierten Systemen noch nicht zugestanden hat, ist die Entwicklung normativer Grundlagen und ethischer Prinzipien. Dabei gehören ethische Entscheidungsmuster zweifellos nicht zu den menschlichen Spitzenleistungen, sondern sie sind eigentlich immer als problematisch und unzuverlässig angesehen worden, so problematisch, dass Luhmann sich immerhin genötigt sah, vor der Moral zu warnen. Gerade in Zeiten grassierender Populismen und Fundamentalismen ist deutlich geworden, dass soziale Systeme gegenwärtig eher unter zu viel Moral als unter zu wenig Moral zu leiden scheinen. Umgekehrt ist ebenso klar, dass soziale Systeme ohne normative Leitplanken nicht funktionieren können, da es ihnen dann nicht gelingt, eine identifizierbare Identität auszubilden.

Wenn also dieses ziemlich zweifelhafte menschliche Residuum, das zumindest als philosophisches Projekt sicher gescheitert ist, auch wenn das die professionelle Philosophie nicht wahrhaben will, durch automatisierte Systeme quasi final herausgefordert wird, dann gibt es offenbar nicht unerheblichen Klärungsbedarf.

Die Technikphilosophie Ernst Kapps (1877, S. 29ff.) ging einst in ihrer These von der Organprojektion davon aus, dass der Mensch sich in den von ihm entwickelten Technologien selbst erkennen könne, und der Medienmythologe McLuhan ordnete bereits in den 60er Jahren des letzten Jahrhunderts den Computer in diese Reihe externalisierter oder ausgelagerter menschlicher Leistungen ein. Die Frage, die sich dann heute stellt, ist, ob das immer noch so ist, denn die neue Konstellation verspräche zumindest in zwei Aspekten neue Einsichten: über automatisierte Systeme und über das menschliche Subjekt selbst.

Erlösende Autonomie

Nicht nur in der Kantschen Vision von Aufklärung, sondern gerade auch in den medientechnologisch stimulierten Prognosen der 60er Jahre des letzten Jahrhunderts wird die Autonomie des Subjekts als Erlösung modelliert, bei Kant noch einigermaßen schwer erkämpft, bei McLuhan und Max Bense eher en passant als eine Art Abfallprodukt der

Medien- und Computertechnologie. Kapps Organprojektion[1], Freuds Prothesengott[2] und McLuhans Erweiterungstheorem[3] dachten Technologien immer vom Subjekt aus und sie dachten sie vor allem als eine Art imperiale Erweiterung des Subjekts selbst und seiner Vermögen, seiner Erkenntnis, seiner Metaphysik, seiner Ästhetik und seiner Handlungsmacht. Zwar mögen Prothesen in der Regel eher als defizitär konnotiert sein, was sie im Übrigen dann mit den Medien gemein hätten, die doch bestenfalls ein schlechtes Surrogat für die eigentliche Kommunikation, den Dialog, zustande gebracht haben, aber ausgerechnet diese Prothesen, die Handlungsmacht, Ästhetik und Erkenntnis versprechen, drohen nunmehr die Macht zu übernehmen und zwar in genau den Regionen, die sie einstmals dem Menschen eröffneten. Aber bekanntlich schlägt das Imperium zurück und die Frage ist, ob es das dann nach den Regeln der Hollywood-Dramaturgie tut und mit welchen Normen Protagonisten und Antagonisten ausgestattet sein werden.

Wenn aber die seit Jahrhunderten versprochene technische Erlösung schon in universelle Entmachtung umzuschlagen droht, dann stellen sich zwei Fragen, die nach der Technologie, ihren strukturellen Implikationen und ihrer inhärenten Normativität sowie die nach der Restgröße des bürgerlichen Subjekts. Kant modellierte das bürgerliche autonome Subjekt als Inkarnation der Vernunft, sämtliche anderen Beigaben: Interessen, Emotionen, Macht und Ökonomie waren bestenfalls störend. Das autonome Subjekt war also eine reine Repräsentation der Vernunft und seine Moral fiel naturgemäß kaum minder vernünftig aus:

---

[1] Vgl. Kapps These der Organprojektion, wonach sämtliche Technologien nichts anderes als Projektionen der Organe des Menschen seien: „Dieses Zustandekommen von Mechanismen nach organischem Vorbilde, sowie das Verständnis des Organismus mittels mechanischer Vorrichtungen, und überhaupt die Durchführung des als Organprojection aufgestellten Princips für die, nur auf diesem Wege mögliche, Erreichung des Zieles der menschlichen Thätigkeit, ist der eigentliche Inhalt dieser Bogen.“ (Kapp 1877, S. VI), und seine Vorstellung, dass der Mensch sich in der Technologie selbst erkennen könne: „Von den ersten rohen Werkzeugen, geeignet die Kraft und Geschicklichkeit der Hand im Verbinden und Trennen materieller Stoffe zu steigern, bis zu dem mannigfaltigst ausgebildeten „System der Bedürfnisse“, wie es eine Weltausstellung gedrängt vorführt, sieht und erkennt der Mensch in all diesen Aussendingen, im Unterschiede von den unveränderten Naturobjecten, Gebild der Menschenhand, Thaten des Menschengeistes, den sowohl unbewusst findenden, wie bewusst erfindenden Menschen – Sich selbst.“ (Kapp 1877, S. 25) fanden in der frühen Medientheorie geneigte Abnehmer.

[2] „Der Mensch ist sozusagen ein Prothesengott geworden, recht großartig, wenn er alle seine Hilfsorgane anlegt, aber sie sind nicht mit ihm verwachsen und machen ihm noch gelegentlich viel zu schaffen. Er hat übrigens ein Recht, sich damit zu trösten, daß diese Entwicklung nicht gerade mit dem Jahr 1930 A.D. abgeschlossen sein wird. Ferne Zeiten werden neue, wahrscheinlich unvorstellbar große Fortschritte auf diesem Gebiete der Kultur mit sich bringen, die Gottähnlichkeit noch weiter steigern“ (Freud 1930, S. 87).

[3] Bekanntlich ist eine von McLuhans im Übrigen keineswegs kohärent zusammen passende Definitionen jene „extensions of man“ – „…alle Medien als Ausweitungen unserer Person dienen dazu, uns neue umformende Einsicht und Bewußtheit zu geben.“ (McLuhan 1964, S. 78) –, die sich zwar an Kapps Konzept der Organprojektion anlehnt, jedoch keineswegs Selbsterkenntnis zum Ziel hat, sondern vornehmlich eine Erweiterung von Handlungsmacht.

Sie war konsequent als Gesetz modelliert und damit war Verallgemeinerbarkeit gleich mit garantiert.[4] Da aber der kategorische Imperativ zugleich als Regel formuliert ist, nämlich als Regel der Universalisierbarkeit – „handle so, als ob die Maxime deiner Handlung durch deinen Willen zum allgemeinen Naturgesetze werden sollte“ (Kant 1785, S. 51) – sollte er prinzipiell eigentlich auch programmierbar sein. Kant liefert damit zugleich eine recht gute Bedingung für eine Anerkennung als autonomes Subjekt, nämlich die Universalisierungsfähigkeit und damit das Vernunftgemäße seiner Entscheidungen.

Von der Quantifizierbarkeit des Menschen

Die bekannten Differenzierungsverfahren, die Mensch und Maschine voneinander scheiden sollten, operierten in utilitaristischer Manier: Sie waren ausschließlich output-orientiert. Was die jeweiligen Akteure antrieb, regulierte und schaltete war hingegen vollkommen uninteressant. Man erfährt etwas über Wahrnehmung und Identifizierung von Maschinen und Menschen, aber man erfährt nichts über die zugrundeliegenden Operationen. Solange der Rezipient im Turing-Test nur keinen Unterschied feststellen konnte, war die Art des Zustandekommens der Information weitgehend gleichgültig und konnte als menschlich zu den Akten gelegt werden.

Insofern macht es die utilitaristische Perspektive sehr viel einfacher als Kants kategorischer Rigorismus mit Turings Probe auf die Maschine (vgl. Turing 1950, S. 39 ff.) zurechtzukommen. Mills Leitidee des größten Glücks der größten Zahl[5] hat den unbestreitbaren Vorzug der Quantifizierbarkeit. Dabei steht in der Mehrzahl der Entscheidungsfälle gar nicht Mills normatives Telos des allgemeinen Glücks zur Disposition. Dieses allgemeine Glück verfügt über einen spekulativen Überschuss, der seine Berechenbarkeit erheblich erschweren dürfte. Aber in der Regel ist das allgemeine Glück in seiner Absolutheit gar nicht gefragt, sondern eine relative Bestimmung reicht vollkommen aus und die zeichnet sich bei Mill durch eine ziemlich problemlose Berechenbarkeit aus.

> „Fragt man mich nun, was ich meine, wenn ich von der unterschiedlichen Qualität von Freuden spreche, und was eine Freude [...] wertvoller als eine andere macht, so gibt es nur eine mögliche Antwort: von zwei Freuden ist diejenige wünschenswerter, die von allen oder nahezu allen, die beide erfahren haben [...] entschieden bevorzugt wird.“ (Mill 1861, S. 15 f.)

Derartige Abstimmungsverfahren, wie sie hier von Mill insinuiert werden, sind ziemlich umstandslos in Entscheidungsverfahren umzusetzen, ja sie lassen sich nicht nur antizipieren und dann als durch Abstimmung gewonnene normative Entscheidungsregeln in Hand-

---

[4] Kant legt bekanntlich einen naturwissenschaftlichen und keinen normativen Gesetzesbegriff zugrunde und orientiert sich damit an dem Allgemeingültigkeitspostulat der Naturwissenschaften (vgl. Kant 1785, S. 51).

[5] „… denn die Norm des Utilitarismus ist nicht das größte Glück des Handelnden selbst, sondern das größte Glück insgesamt.“ (Mill 1861, S. 20)

lungsroutinen autonomer Systeme implementieren, sie basieren eben auch in letzter Hinsicht auf Entscheidungen des autonomen Subjekts. Zwar braucht Mill entschieden mehr Personal als Kant[6], um zu seinen normativen Entscheidungen zu gelangen. Allerdings ist das in Zeiten von Big Data auch kein ernstliches Problem. Das Entscheidende ist, dass Mill die absoluten normativen Entscheidungen Kants auf beobachtbare Abstimmungsergebnisse heruntertransformiert und damit die Beherrschbarkeit durch Berechenbarkeit garantiert. Bei Kant wäre das zumindest nicht ganz so einfach. Zwar vergisst auch Mill bei seinen berechenbaren Entscheidungen nicht die Ansprüche der Viktorianischen Klassengesellschaft, die nicht unbedingt für ihre Liberalität und einen Mangel an Klassenbewusstsein bekannt war, zu wahren, indem er für die Teilnahme an einer Entscheidung die praktische Erfahrung der Folgen voraussetzt und damit den Kreis der Entscheidungsberechtigten auf ein ‚erträgliches' Maß zurechtstutzt und so die Klassenintegrität[7] wahrt, aber gegenwärtig wäre das Ganze durchaus auch unter demokratisch akzeptableren Konditionen denkbar. Es gibt also kein prinzipielles Hindernis gegen ein utilitaristisches Entscheidungsverfahren.

Das Problem ist nur, dass utilitaristische Ansätze prinzipiell ex post steuern und das ist bei Technologien mit einem enormen Erfassungsradius und einer ebensolchen gesellschaftlichen Relevanz alles andere als unproblematisch. Vor allem ergibt sich bei komplexen Technologien die Schwierigkeit, dass eine Folgenabschätzung kaum mit hinreichender Zuverlässigkeit durchzuführen ist. Sowohl die Neuheit als auch der enorme Einzugsbereich digitaler Technologien erschweren daher eine normative Prognose und erhöhen damit zugleich das Risiko einer Steuerung ex post. Die Simplizität und prinzipielle Quantifizierbarkeit des utilitaristischen Ansatzes werden mithin mit dem prinzipiellen Zuspätkommen des Steuerungsimpulses erkauft. Utilitaristische Steuerungslogiken müssen also die Reversibilität der fraglichen Handlungen voraussetzen, andernfalls handelte es sich um eine ziemlich riskante Experimentalanordnung.

Für die üblichen sozialen Prozesse und Interaktionen mag aufgrund einer einigermaßen zuverlässigen Prognose der möglichen gesellschaftlichen Risiken eine solche Steuerung ex post noch mit einigermaßen überschaubaren Folgen verbunden sein und daher durchaus als akzeptabel erscheinen. Für neue Technologien hingegen gelten diese Routinen allenfalls in einem sehr viel eingeschränkteren Maße. Zwar lassen sich gerade auch für die

---

6 Kant benötigte bekanntlich nur ein, quasi als paradigmatisch gesetztes Subjekt; Sozialsysteme hingegen mit ihren Ideologien, kulturellen Lasten und Traditionen waren für ihn vollkommen irrelevant. Moralisch gesehen störten sie eigentlich nur, was, da es bekanntlich um soziale Entscheidungsgründe geht, letztlich einen Widerspruch darstellt. Man hat es also bei der Kantschen Konstruktion mit einer Art moralischer Robinsonade zu tun, die sich ihren Freitag allenfalls als eine Art Spiegelung denkt.

7 Für eine solche politische Klugheit bzw. für einen solchen Opportunismus oder ‚Realismus' ist Mill ja bekannt. So scheut er sich auch nicht, den Sklavenhandel, also einen Grundbestandteil der britischen Dreiecksökonomie, normativ zu legitimieren. (vgl. Mill 1861, S. 55)

Implementation neuer Technologien, insbesondere aber für die von Kommunikationstechnologien morphologische Strukturmuster – also gesellschaftliche Routinen – feststellen, jedoch ist vor allem bei Kommunikationsroutinen die Abschätzung möglicher Folgen äußerst vage.

Daher operieren Gesellschaften in diesem Fall auch mit einer Art Intervallschachtelung zwischen Apokalypse und Erlösung[8] und können sich dann überlegen, welchen Einsatz sie gesellschaftlich letztlich riskieren wollen. Die Verengung und Justierung des Spektrums zwischen Apokalypse und Erlösung ist eine Angelegenheit, die nur mittels gesellschaftlicher Kontrolle einigermaßen zuverlässig stabilisiert werden kann. Gesellschaftliche Kontrolle und Steuerung aber funktioniert zunächst einmal normativ durch Errichtung eines durch Normen bewährten Cordon Sanitaire, in dem die Technik sich tummeln darf. Wenn aber bei neuen Technologien und vor allem bei generellen Umbrüchen technologischer Paradigmen wie der Digitalisierung die Abschätzung der gesellschaftlichen Folgen mit erheblicher Unzuverlässigkeit belastet ist, dann basiert die Festlegung eines solchen Cordon Sanitaire auf kaum mehr als einer bloßen Mutmaßung, einer Mutmaßung, die immerhin nicht weniger die Kommunikationsstrukturen von sozialen Systemen regulieren soll.

Eine Steuerung ex post muss daher mit einem nicht unerheblichen Schadensaufkommen rechnen. Zugleich wird deutlich, dass wir uns akut faktisch in eben dieser Situation eher erfolgloser nachgelagerter Steuerungsversuche befinden. Umgekehrt erscheinen nachgelagerte normative Verfahren immer dann einigermaßen unproblematisch zu sein, wenn der Umfang der technischen Innovation als halbwegs überschaubar erachtet wird oder zumindest als überschaubar dargestellt werden kann. Dabei geht es weniger um die faktische Wirkmächtigkeit von Technologien als um die symbolische Differenz, mit der sie kommuniziert werden. Sobald sich metaphorische Brücken bauen lassen und Plausibilisierungen qua Analogie zu verfangen scheinen, dann fällt die symbolische Differenz relativ gering aus und nachgelagerte Steuerungslogiken werden problemlos akzeptiert.[9] Auf diesem

[8] Technische Innovationen werden, sobald sie auch nur ansatzweise über gesellschaftliche oder kulturelle Relevanz verfügen, von den betroffenen Interessengruppen regelmäßig kontrovers diskutiert: Die Verfechter der Technologie treten üblicherweise mit der Vorhersage der Heilung aller aktuellen gesellschaftlichen Mängel also mit einem Erlösungsversprechen an, diejenigen, die sich von der neuen Technik bedroht sehen, prophezeien hingegen den kulturellen und sozialen Untergang und modellieren die neue Technologie in Termini des Verlusts. Aus dieser Dialektik von Apokalypse und Erlösung entwickelt sich im Zuge der Implementation der neuen Technologie dann eine domestizierte Technologie, der spezifische kulturelle Beschränkungen auferlegt werden und die dann als solche enkulturiert wird.

[9] Genau das aber ist im Fall autonomer Fahrzeuge das Problem: Assistenzsysteme, wie kompetent sie auch immer sein mögen, werden als relative und daher überschaubare technische Erweiterungen verstanden. Sie stellen insofern auch keine prinzipiellen Fragen, wie weitreichend diese Systeme auch immer angelegt sein mögen. Die behauptete Autonomie hingegen stellt die Prinzipienfrage und steigert die Folgen notwendig ins Unermessliche. Wenn autonome Systeme Fehler machen oder versagen, dann handelt es sich um etwas vollkommen anderes, als wenn ein Assistenzsystem eine Fehl-

Wege können Technologien quasi schleichend eingeführt und normativ plausibilisiert werden. Hierbei sorgt schlichte Gewöhnung und die Niedrigschwelligkeit des Verfahrens für ein trügerisches Vertrauen in Technologien. Die nachgelagerte Steuerung eignet sich mithin vor allem für relativ geringe Innovationsniveaus, die Innovationsfrequenz ist demgegenüber von eher untergeordneter Bedeutung, weil der Prozess nahezu beliebig oft wiederholbar ist.

Das Verfahren der kumulativen normativen Enkulturation von Technologien eignet sich jedoch aus systematischen Gründen nicht bei gravierenden technologischen Umbrüchen oder technologischen Veränderungen mit hoher Innovationstiefe. Bei derartigen technologischen Paradigmenwechseln versagen Analogien als Argumentationshilfe, da deren Unangemessenheit unmittelbar einleuchtet. Das Neue wird allein negativ bestimmbar und gewinnt dementsprechend keine Gestalt. Die sozio-kulturellen und ökonomischen Folgen von solchen Paradigmenwechseln lassen sich ebenso wenig verlässlich bestimmen. Die Idee auch solche technologischen Paradigmenwechsel noch durch nachgelagerte Steuerung in den Griff bekommen zu können, operiert vergleichbar jener ebenso prominenten wie eklatant gescheiterten Devise des Kölschen Klüngels „Es ist noch immer alles gut gegangen". Derart seliger Optimismus erscheint angesichts des sozio-kulturellen und ökonomischen Impacts gravierender technologischer Umbrüche zusehends naiv.

Vom Turing-Test zur Prinzipienfrage oder: Zurück zur Autonomie

Kants Vorstellung vom Subjekt geht jedoch sowohl über die utilitaristische Quantifizierungsstrategie als auch über den Turing-Test weit hinaus, indem er nicht nur die einigermaßen beliebige Wahrnehmbarkeit als Subjekt zur Bedingung macht, sondern den kategorischen Imperativ zugleich als eine Art Testfrage formuliert, an der sich mögliche Differenzen und damit die Unterscheidung von Prothesen und Subjekten erweisen könnten.

Der deontologische Ansatz Kants geht vollkommen anders vor, indem er nicht nur von einer a priori ansetzenden Steuerung, also einer vorgelagerten normativen Konstruktion ausgeht, sondern indem er den automatisierbaren Akteur selbst zu bestimmen sucht. Dabei wird zugleich deutlich, dass Kant in seiner rigorosen Nüchternheit eigentlich eine Maschinenethik[10] geschrieben hat. Der kategorische Imperativ Kants diskriminiert bekanntlich

---

funktion aufweist, selbst wenn der Schaden identisch sein sollte: Im ersten Fall stellt sich die Prinzipienfrage und die verlangt in jedem Fall nach kategorischen Antworten, im zweiten Fall stellt sich einerseits die Frage nach technischen Lösungen, die den Fehler künftig verhindern könnten, und andererseits die Frage nach der Verantwortung für den technischen Fehler. Die gesellschaftliche Reaktion unterscheidet sich dann demgemäß auch zwischen Verbot und Verbesserung.

[10] Das scheint schon Schiller geahnt zu haben, wenn er das ‚Technische' bzw. die ‚technische Form' der Vernunft bei Kant moniert: „Über diejenigen Ideen, welche in dem praktischen Teil des Kantischen Systems die herrschenden sind, sind nur die Philosophen entzweit, aber die Menschen, ich getraue mir, es zu beweisen, von jeher einig gewesen. Man befreie sie von ihrer technischen Form und sie werden als die verjährten Aussprüche der gemeinen Vernunft und als Tatsachen des moralischen Ins-

anhand der Universalisierbarkeit von Entscheidungen. Auf dieser Basis lässt sich auch zwischen autonomen und nicht-autonomen Systemen unterscheiden, indem eben diese Universalisierungsfähigkeit zum Differenzkriterium gemacht wird. Der Kant-Test, der dementsprechend über die Autonomie von was auch immer entscheidet, lautete dann: Erst wenn eine Person oder ein System in der Lage ist, seine Entscheidungen zu universalisieren, handelt es sich um ein Subjekt oder um ein autonomes System. Der Kant-Test, der quasi die Vernunftfähigkeit des Subjekts oder des automatischen Systems beurteilt, unterscheidet zwischen Systemen, die autonom, also ohne Kontrolle agieren können, und solchen, die in jedem Fall einer Kontrolle durch ein autonomes Subjekt bedürfen. Systeme, die ohne prinzipielle Vorabkontrolle auskommen, sind entweder solche, die über überschaubare Innovationsniveaus verfügen und sich von bereits enkulturierten Technologien nicht wesentlich unterscheiden, oder aber Technologien, die ohnehin über explizite Kontrollen verfügen. Solche Systeme sind entweder prinzipiell nicht subjektfähig oder aber sie werden ohnehin Subjekten unterstellt. Verantwortung wird in einem solchen Fall dem kontrollierenden Subjekt oder aber dem Produzenten der Technologie zugewiesen.

Für solche abhängigen Systeme gilt das technische Sklavenmodell: Technologien werden Subjekten als vollkommen abhängige Unterstützungsinstrumente zugeordnet und all ihre Operationen sind letztlich von den Entscheidungen dieses Subjekts abhängig. Das heißt, dass diese Systeme jederzeit durch Subjekte abschaltbar sein müssen und das Subjekt prinzipiell die Verantwortung für den Einsatz der Systeme trägt.

Alle anderen Systeme, die also nicht irgendwelchen Modi der Nachkontrolle unterliegen sollen, müssen den Kantschen Subjekt-Test bestehen, d.h., sie müssen eine Selbststeuerungsfähigkeit auf Universalisierungsniveau unter Beweis stellen. Gleichzeitig würde mit einem in diesem Sinne erweiterten Turing-Test aber auch die Perspektive gewechselt: Man würde von der Rolle des unbeteiligten Betrachters von Handlungen und Handlungsfolgen zur Analyse und Steuerung der Motivation von Handelnden übergehen. Die Anerkennung von Systemen als Subjekte griffe quasi a priori in die Konstruktion des Systems ein, indem sie eine Implementation einer Universalisierungsroutine in jeden Entscheidungsprozess erforderlich machte, die diesem System zu Grunde läge.

Die Universalisierungsfähigkeit stellt zugleich eine relativ große Hürde dar, da sie prinzipiell nicht durch irgendwelche Quantifizierungen zu substituieren ist. Denn selbst wenn man Schadensminimierung als verallgemeinerungsfähiges Ziel akzeptierte, funktionierte das allenfalls noch bei der Schädigung von Objekten, bei der Schädigung von mensch-

tinktes erscheinen, den die weise Natur dem Menschen zum Vormund setzte, bis die helle Einsicht ihn mündig macht. Aber eben diese technische Form, welche die Wahrheit dem Verstand versichtbart, verbirgt sie wieder dem Gefühl; …“ (Schiller 1795, 1. Br., S. 4) Schiller geht es bekanntlich darum, „den Gesetzen der Vernunft ihre moralische Nötigung“ (Schiller 1795, 14. Br., S. 58) zu nehmen und d.h. die Rigidität des Gesetzes ‚erträglich‘ zu machen. Dieses Problem stellt sich bei autonomen Systemen jedoch gerade nicht.

lichen Individuen jedoch wäre Quantifizierung systematisch unangemessen. Insofern dürften, folgt man dem Kantschen Denkmodell, Maschinen, die prinzipiell nicht in der Lage sind, menschliche Individuen zu schädigen und die gleichzeitig nicht über die erforderlichen Universalisierungsfähigkeiten verfügen, nicht als autonome Systeme in den Verkehr gebracht werden. Das macht zugleich die vergleichsweise hohe Hürde deutlich, die beim Übergang von teilautonomen zu autonomen Systemen besteht.

Verfahren normativer Kontrolle

Selbst prinzipiell universalisierungsfähige Subjekte unterliegen der Kontrolle. Kaum ein gesellschaftliches System traut der Selbststeuerungsfähigkeit von Subjekten wirklich, sondern sie operieren allesamt mit mehr oder minder ausgefeilten Polizei- und Justizsystemen, die dem nachlässigen Subjekt auf die Sprünge helfen sollen. Je größer die sozialen Differenzen in Sozialsystemen sind, umso größer wird zwangsläufig die zur Aufrechterhaltung des Ungleichgewichts erforderliche Kontrolle. Der Raum der Autonomie von souveränen Subjekten ist daher stets mehr oder minder deutlich eingeschränkt. Die Souveränität von autonomen Subjekten wird so durch soziale Verfahren der Kontrolle eingehegt, was umgekehrt bedeutet, dass die Autonomie des Subjekts nichts anderes als ein regulatives Prinzip darstellt. Bei solchen regulativen Prinzipien wird davon ausgegangen, dass sie in der Praxis nie allein oder ausschließlich wirksam werden können, sondern stets von anderen Interessen überlagert zu werden drohen. Derartige Interferenzen fallen aber bei autonomen Systemen systematisch weg oder aber sie werden durch technische Störungen bzw. ökonomische Interessen abgelöst. Insofern verhalten sich autonome Systeme zu regulativen Prinzipien wie der Idee der Universalisierungsfähigkeit von Entscheidungen zwar grundlegend anders als autonome Subjekte, jedoch nicht notwendig auch optimaler.

Wenn aber allen anderen Systemen, die nicht die Universalisierungsfähigkeit in ihren Entscheidungsroutinen implementiert haben, normative Souveränität und damit selbst die störungsanfällige Selbststeuerung durch regulative Prinzipien fehlt, dann müssen sie zwangsläufig vollständig gesellschaftlicher Kontrolle unterworfen werden. Die Modi einer solchen gesellschaftlichen Kontrolle unterscheiden sich dabei durchaus: Das enkulturierte Standardmodell für die gesellschaftliche Steuerung von Technologien besteht in einem zweistufigen Steuerungsmodell: Selbstkontrolle durch das Artefakt produzierende oder betreibende Subjekt sowie einen institutionalisierten Kontrollmechanismus, der in der Regel von öffentlich-rechtlichen oder aber staatlichen Institutionen organisiert, durchgeführt oder beauftragt wird. Das Ganze wird dann im Streitfall noch einer juristischen Nachkontrolle unterzogen, sodass man es letztlich mit einem dreistufigen Kontrollmodell zu tun hat. Die Medienkontrolle funktioniert in der Regel nach solchen dreistufigen Steuerungslogiken.

Das aber würde bedeuten, dass sämtliche teilautonomen Systeme solchen wenigstens einigermaßen gesellschaftlich anerkannten Steuerungslogiken unterworfen werden müssten und der Streit, der gegenwärtig um solche technischen Systeme herrscht, letztlich vor al-

lem damit zu tun hat, dass derartige Steuerungsmechanismen noch nicht etabliert sind. Seit John Perry Barlows einigermaßen naiver Unabhängigkeitserklärung des Cyberspace von 1996 wird im Internet mit dem regulativen Prinzip der Autonomie und der Selbstregulierung operiert, was letztlich darauf abzielt, möglichst alle Formen gesellschaftlicher Kontrolle und sekundärer Steuerung auszuschließen. Das hinzu kommende eklatante Versagen juristischer Steuerungsstrategien generierte einen normativ weitgehend unregulierten Raum, in dem keineswegs individuelle Kreativität wucherte, sondern ökonomische Interessen die regulativen Prinzipien der Selbstkontrolle und Autonomie behänd außer Kraft setzten. Die Diskussion um die Geschäftsmodelle von sozialen Netzwerken und Suchmaschinen hängt nicht zuletzt mit der eklatanten Unwirksamkeit der Selbstregulation und der Verdrängung regulativer Prinzipien durch ökonomische Interessen Einzelner zusammen.

Derartige Konflikte sind für den Enkulturationsprozess von neuen Medientechnologien geradezu typisch. So wurden parallel zu ihrer Etablierung sowohl der Buchmarkt als auch der Rundfunk Regulationsstrategien vom Urheberrecht bis zur staatlichen Erteilung und Kontrolle von Funklizenzen unterworfen. Offenkundig ist die Enkulturation von Internet und autonomen Systemen noch nicht so weit fortgeschritten, dass solche Regulationsstrategien entwickelt und durchgesetzt worden wären. Dass die Privilegierung ökonomischer Interessen Einzelner zu normativen Fehlentwicklungen führt, dürfte vermutlich den meisten Akteuren zwischenzeitlich deutlich geworden sein.

Haltet die Betrüger! Oder die Strategien der Moralisierung

Social Bots bestehen, so viel scheint einigermaßen sicher zu sein, den Kant-Test nicht und d.h., sie sind in jedem Fall einer wie auch immer gearteten sozialen Kontrolle zu unterstellen. Nur ist Voraussetzung ihres Erfolgs, dass sie es verstehen, sich der Kontrolle zu entziehen. Dabei beruht das Prinzip ihrer Wirkung auf der Imitation autonomer Subjekte ohne selbst autonom zu sein. Vorgebliche Souveränität, hinter der sich nicht offen gelegte Interessen verbergen, wird, sofern sie ihrer Vorgeblichkeit überführt wird, als ein Phänomen der Täuschung begriffen und so wird der Diskurs über Social Bots vornehmlich von Betrugsnarrativen regiert.

Dieses Betrugsmotiv ahnte bereits Turing, denn er benutzte es bei der Klärung der Frage „Können Maschinen denken?“ (Turing 1950, S. 39). Er transformierte nämlich die Frage in ein Imitationsspiel[11] und Imitation ist bekanntlich nichts anderes als ein Spiel mit Täu-

[11] Turing formuliert die Frage „Kann eine Maschine denken?“ um in ein Imitationsspiel: „Die neue Form des Problems läßt sich als Spiel beschreiben, das wir ‚Imitationsspiel‘ nennen wollen. Wir betrachten drei Spieler, einen Mann (A), eine Frau (B) und einen männlichen oder weiblichen Fragesteller (C). Der Fragesteller sei allein in einem Raum. Das Ziel des Fragestellers ist es zu entscheiden, welche der beiden anderen Personen der Mann bzw. die Frau ist. […] Wir stellen nun die Frage: Was passiert, wenn eine Maschine die Rolle von A in diesem Spiel übernimmt? […] Diese Fragen treten an die Stelle unserer ursprünglichen: ‚Können Maschinen denken?‘“ (Turing 1950, S. 39f.)

schungen. Die Logik dieses Imitationsspiels besteht darin, dass ein nicht entdeckter Unterschied[12] als Gleichheit, zumindest aber als Gleichwertigkeit akzeptiert wird. Das Wirkungsprinzip von Social Bots basiert auf solch einer angenommenen Gleichwertigkeit und auf den entsprechend unentdeckten Unterschieden. Social Bots haben den Turing-Test, der in diesem Imitationsspiel besteht, immer schon bestanden, wenigstens, wenn sie nicht entdeckt wurden. Umgekehrt bestände, wenn Social Bots entdeckt würden, ja auch keine Gefahr mehr: die Täuschung wurde entlarvt und die Angelegenheit stellt sich mithin als harmlos heraus. Das Bestehen des Turing-Tests ist also dasjenige, was Social Bots letztlich gefährlich macht.

Täuschungs- und Manipulationsvorwürfe, Überwältigungsvorstellungen und Ängste vor einem drohenden Kontrollverlust befeuern dabei die moralischen Diskurse und diese moralischen Diskurse sind letztlich nur Ausdruck eines Verlusts an Unterscheidungskraft. Es geht bei dem Ganzen um die Aufrechterhaltung einer Grenze, die an sich nicht mehr feststellbar ist. Moralisierung tritt immer dann auf den Plan, wenn der Sachgrund für eine Unterscheidung weggefallen ist. Die Diskussion kann also nur noch moralisch geführt werden, denn die Sache selbst wird eigentlich bereits verloren gegeben. Allerdings verfügt dieser an sich so unmittelbar einleuchtend erscheinende Diskurs selbst über mindestens zwei blinde Flecken und das lässt ihn selbst unweigerlich zu einem strukturellen Täuschungsdiskurs werden.

Die echauffierten Brandreden der Hüter der traditionellen Diskurse gehen davon aus, dass Social Bots die Diskurse, Daten und Informationen, die sie lancieren, lostreten und zur Verfügung stellen, unrechtmäßig verfälschen, verzerren oder aber konterkarieren. Eingeübte Kommunikationspraktiken verfangen nicht mehr in der gewohnten Weise oder sie sind schlicht nicht mehr verlässlich. Der Blick auf jene berechenbaren Rezipienten, und an eben dieser Berechenbarkeit haben persuasive Kommunikationsstrategen seit langem mit aller Kraft gearbeitet, droht aufgrund der unkalkulierbaren Intervention diverser Social Bots eine Art grauen Star zu bekommen und zusehends wieder schleierhaft zu werden. Der gerade eben erst nahezu vollständig kalkulierbar gewordene Rezipient droht in einer Wolke von Bots und traditionellen Normalbürgern wieder in traditioneller Unsicherheit zu verschwinden. Social Bots produzieren weitgehend unbeabsichtigt einen enormen Kollateralschaden, denn sie greifen völlig unabhängig von dem Kalkül ihrer Konstrukteure zunächst einmal das Kapital der Datensammler und der Eigentümer der digitalen Netzwerke an. Es wird mithin ein recht erfolgreiches Geschäftsmodell destabilisiert und das gibt natürlich Ärger. Die Akteure mit Definitionsmacht in den Diskursen und die Datensammler wehren sich und sie wehren sich, solange sie juristisch an die Konstrukteure der durch die Netzwerke nomadisierenden Bots nicht herankommen können, durch moralische Verurteilung. Und bei all diesen Brandreden geben sie vor, alles nur im Dienste jener vom Radar wieder zu verschwinden drohenden Rezipienten zu tun.

---

[12] In Turings Fall der von Mensch und Maschine.

Nun, wenn eines in der aufgeregten Diskussion um Social Bots sicher ist, dann das, dass der moralisierende Diskurs der Akteure mit Definitionsmacht im Wesentlichen selbst auf einer Täuschung beruht. Es ist nämlich durchaus im Interesse des Rezipienten, wenn seine ubiquitäre Erfassbarkeit zusehends Risse bekommt. Und die Erkenntnis, dass der Rezipient den Informationen der Akteure nicht mehr trauen kann, ist ein durchaus probates Erziehungsmittel[13], das Resultate zeitigen könnte, für die die kritische Theorie noch vor nicht allzu langer Zeit erhebliche theoretische Klimmzüge machen musste. Misstrauen gegenüber den Protagonisten der Persuasion ist also angebracht und es ist normativ gesehen ziemlich egal, ob diese sich echauffieren oder nicht. Der ganze moralisierende Diskurs sucht also den Rezipienten vor den Karren von Datenverkäufern, Meinungsmachern, Trendsettern und politischen Spindoctoren zu spannen und es ist zweifellos gut, wenn ein solches Vorhaben scheitert. Von daher haben Social Bots eine ganz sicher nicht intendierte subversive Nebenwirkung[14], deren Effekt vor allem in der Streuung von Misstrauen besteht.

Manipulationsvorwürfe im Zusammenhang von Medien sind im Übrigen alles andere als neu, und der Verdacht war immer schon so hinreichend begründet, dass gar nicht erst irgendwelche Verschwörungsszenarien bemüht werden mussten. Die Akteure mit Definitionsmacht innerhalb des Mediensystems, und das schließt auch die digitalen Medien ein, blicken auf eine lange Tradition des Kampfes gegen eben jenen Manipulationsverdacht zurück: Das von Journalistenverbänden und Verlegern gleichermaßen fetischisierte Objek-

---

[13] So wittert die Medienpädagogik auch hier wieder – wie bei nahezu jedem Medienskandal – Morgenluft und bietet Media Literacy als probates Gegenmittel für die Täuschungsstrategien von Social Bots und disruptiver Kommunikation: „This development strikes at the heart of democratic societies in the western world, which rely on communication in regards to democratic processes, particularly general elections and political debates. It is against this background that social bots and other forms and manifestations of disruptive communications, like fake news, have to be understood" (Janetzko 2017, S. 65 f.). Die Medienpädagogik bietet sich hier ein weiteres Mal als Reparaturbetrieb für die Fehlfunktionen des Mediensystems an, trägt jedoch nichts zur grundlegenden Lösung der Widersprüche bei.

[14] Auf solche nicht-intendierte Nebeneffekte setzt dann auch Leistert mit seinem sozialromantisch eingefärbten Piraterieszenario: ‚Dieser vermeintlich parasitäre Charakter von Piraterie verkennt jedoch dessen produktive Seite: Piraterie hat über unterschiedliche Märkte verteilt höchst kreative Formen der Distribution erfunden. Und dies ist die entscheidende Parallele, die diese Piraterie mit der Art verbindet, wie Social-Bots Daten weiterverteilen. Der Unterschied ist jedoch der Modus der Distribution. Daten, die Bots gesammelt haben, können in höchst unterschiedliche Richtungen verteilt werden, vom Kreditkartenbetrug bis zum Hacken von Websites, von erstaunlich passenden Werbeeinblendungen oder Spam bis zum Identitätsdiebstahl. Jedoch, […], grundsätzlich sind sie für interessierte Käufer jeglicher Couleur verfügbar, exakt wie die sogenannten ‚Medien der Piraten'" (Leistert 2017, S. 227). „In diesem Sinne ist Piraterie vielmehr als komplementär denn als parasitär zu verstehen, […]" (Leistert 2017, S. 228). „Insofern ist es durchaus möglich, Social Bots jenseits moralischer oder legalistischer Überlegungen in einer der Datenakkumulation verschriebenen kapitalistischen Ökonomie, die selber die Privatsphäre und informatische Selbstbestimmung strukturell angreift, zu situieren" (Leistert 2017, S. 228).

tivitätsideal[15], das nicht nur zwischen Journalismus und PR trennen helfen soll, sondern den Rechtsgrund gleich für ganze Serien von Privilegien vom Zeugnisverweigerungsrecht bis hin zu Neuwagenrabatten bildet, soll eben auch den Manipulationsverdacht abwehren.

Dass ausgerechnet genau dieser Verdacht nun gegen die die Definitionsmacht zu erschüttern drohenden Social Bots gerichtet wird, ist mehr als eine bloße Reflexreaktion. Er setzt nämlich voraus, dass die Akteure selbst prinzipiell gute und das hieße universalisierungsfähige Interessen vertreten. Die Identifikation betrügerischer Social Bots exkulpiert automatisch ausgerechnet diejenigen Akteure, denen es bis dato letztlich nicht nachhaltig gelungen ist, sich selbst des Manipulationsverdachts zu erwehren, also diejenigen, deren Interesse in der möglichst vollständigen Erfassung von Konsumenten besteht. Dass auf diesem Wege ausgerechnet die Besitzer von sozialen Netzwerken und Datensammlungen moralischen Dispens erhalten und auf einmal auf die Seite der bedingungslos Guten wechseln dürfen, kann nur als historische Ironie des Digitalisierungsprozesses gewertet werden.

Würde mit Social Bots nicht gleichzeitig das Geschäftsmodell sozialer Netzwerke[16] bedroht, könnte man auf die Idee kommen, dass den Akteuren mit kultureller Definitionsmacht und mit der Verfügungsgewalt über Datensammlungen nichts Besseres als Social Bots hätte passieren können. Normalerweise kostet der bedingungslose Wechsel auf die Seite der Guten zumindest normative Unterwerfung. Dazu wurden die Sozialen Netzwerke noch nicht einmal genötigt, stattdessen fallen durch die Erosion des Geschäftsmodells nun ökonomische Kosten[17] an.

Das gilt letztlich auch für das Geschäftsmodell der empirischen Sozialwissenschaften, sofern sie versuchen, soziale Experimente unter Verwendung von Social Bots als neue Chance sozialwissenschaftlicher Forschung zu verkaufen. Social Bots drohen sozialwissenschaftliche Analysen öffentlicher Kommunikation allein durch schlichte Verunsiche-

---

[15] Aus dieser Perspektive wird eben auch ein Großteil der Fake News Debatte und der moralisierenden Diskussion um Social Bots geführt, so dass es sich hierbei auch um eine der so häufigen Auseinandersetzung um mediale Definitionsmacht geht. Der traditionelle Journalismus sieht sich zunehmend durch Social Media und automatisierte Textproduktion bedroht und versucht normativ nachzusteuern, um seine privilegierte Stellung zu erhalten.

[16] „Social Bots stellen langfristig eine Bedrohung für das Geschäftsmodell von sozialen Netzwerken dar. Ein Teil der Nutzer könnte sich abwenden, weil sie das Vertrauen in die Echtheit der Beiträge verlieren. Investoren verlieren das Interesse, weil sich die Plattformen durch Werbeeinnahmen oder dem Verkauf von Nutzerdaten finanzieren, aber nur echte Menschen Kaufentscheidungen treffen.“ (Kind, Bovenschulte, Ehrenberg-Silies u.a. 2017, S. 5)

[17] Eine ebenso de-moralisierende wie ganzheitliche Perspektive nimmt Leistert ein: „Dieser Beitrag nimmt deshalb einen anderen Weg, und versucht Bots aus der Perspektive ihres Evironments zu untersuchen, als Teil und Element einer medien-technischen Umgebung. Gleichzeitig ist es notwendig, Bots ins Verhältnis zur Logik des zeitgenössischen kapitalistischen Imperativs der Datenextraktion und dem Kolonisieren jeglicher, noch so unbedeutender Äußerungen von Benutzern auf den Plattformen zu setzen“ (Leistert 2017, S. 215).

rung zu kontaminieren und sie konterkarieren damit eben jene Vorstellung, es ließe sich mithilfe von Social Bots ein unverstellter Zugang zu öffentlichen Kommunikationsstrukturen gewinnen. Zugleich droht der harsch moralisierende Diskurs zurückzuschlagen: Täuschung als Instrument einer wie auch immer vermittelt, der ‚Wahrheit' verpflichteten Wissenschaft scheint zumindest das Potential zu haben, sich selbst zu demontieren. Wenigstens aber dürfte die Legitimität einer wesentlich mit Täuschungen operierenden Wissenschaft in Frage stehen. Dass der Einfluss von Social Bots zudem noch als vergleichsweise niedrig veranschlagt wird, nimmt vielleicht ein wenig Dampf aus der Diskussion, ändert an der Dialektik von Apokalypse und Euphorie jedoch kaum etwas.

Der ewige Kampf um Kontrolle

Insofern drohen Social Bots stets das System zu unterlaufen, von dem sie eigentlich zu profitieren gedachten. Täuschung funktioniert nur so lange, solange sie nicht entdeckt oder zumindest nicht erwartet wird. Die Erwartung der Täuschung jedoch lässt das kommunikative System, dem die Social Bots parasitär verpflichtet sind, implodieren. Das Kalkül hinter Social Bots ist daher in jedem Fall ein Kalkül auf Zeit, faktisch sogar auf ziemlich kurze Zeit. Und dieser begrenzte Zeitraum, in dem Social Bots ihre ohnehin vergleichsweise beschränkte Wirksamkeit[18] entfalten, ist das historische Intervall, in dem der Kampf um die Definitionsmacht[19] über neue Medien und Kommunikationskulturen ausgefochten wird.

Diese normativen Auseinandersetzungen sind aber immer schon Kämpfe um die Definitionsmacht in und die Kontrolle über Medien gewesen. Und genau das ist auch bei Social Bots und autonomen Systemen der Fall: Es geht um die Entwicklung und Durchsetzung gesellschaftlich akzeptierter Strategien der Kontrolle. Das normative Risiko von Social Bots besteht nicht so sehr in dem Schaden, den sie unmittelbar, also durch Täuschung

---

18 „Es gibt lediglich eine begrenzte Anzahl prominenter Beispiele der Einflussnahme durch Social Bots, auf die sowohl in der Presse als auch in wissenschaftlichen Artikeln immer wieder Bezug genommen wird. Der in den Artikeln beschriebene Wirkungsraum ist an erster Stelle Twitter und schon seltener Facebook. Das Ausmaß der tatsächlichen Einflussnahme ist allerdings noch kaum belegt" (Kind et al. 2017, S. 5).

19 „Genau an dieser Stelle wird offenbar, warum der Einsatz von Social Bots durch Parteien verfassungsrechtlich bedenklich ist: Er missachtet das demokratische Willensbildungsprinzip des „Von-unten-nach-oben." Mit Hilfe von Social Bots wird in der sozial-medialen Öffentlichkeit zwar der Eindruck erweckt, dass eine Meinung von vielen individuellen Bürgern stammt. In Wahrheit jedoch hat sie ihren einzigen Ausgangspunkt in einer Parteizentrale. Sie ist eine von oben in die Bot-Profile diktierte Meinung" (Gasser & Kraatz 2017). Hier wird deutlich, dass das Öffentlichkeitsmodell des traditionellen Journalismus und seine Logik der politischen Willensbildung sich durch Social Bots unterlaufen sieht. Es handelt sich mithin eben auch um eine Auseinandersetzung um die politische Definitionsmacht zwischen den traditionellen Akteuren des politischen Journalismus und den hinter Social Bots stehenden Akteuren.

anrichten, sondern in den gesellschaftlichen Kontrolldefiziten, über die sie hinwegtäuschen.

Dabei stellen Social Bots und autonome Systeme die gesellschaftliche Kontrolle noch nicht einmal vor besonders große Herausforderungen: Denn bereits an den analogen Medien sind Kontrollstrategien entwickelt und erprobt worden, die sich durchaus auch auf soziale Netzwerke und autonome oder teilautonome Systeme beziehen lassen. Die gegenwärtig einigermaßen aufgeregte und empörte Diskussion dürfte dabei nur die Begleitmusik für den Übergang von privaten, liberal kapitalistisch organisierten Kommunikationssystemen hin zu Systemen darstellen, die einer wie auch immer gearteten gesellschaftlichen Kontrolle unterworfen sind.

Weit und breit findet sich also nichts Besonderes und zudem ist die Angelegenheit vorübergehender Natur. Zu beobachten bleibt jedoch, wer mit welchen Interessen sich an der Auseinandersetzung um die Definition neuer Medien und autonomer Systeme beteiligen wird, aber auch dabei wird es vermutlich kaum Überraschungen geben.

Literatur

Freud, Siegmund (1930): *Das Unbehagen in der Kultur*. Frankfurt a. M.: S. Fischer 379.-408. Tsd. 1970.

Gasser, Lucas & Kraatz, Matthias (2017): Social Bots: Wegbereiter der Maschinokratie. Abgerufen unter: Verfassungs-blog.de/social-bots-wegbereiter-der-maschinokratie [Stand vom 16-01-2017].

Janetzko, Dietmar (2017): Social Bots and Fake News as (not) seen from the Viewpoint of Digital Education Frameworks. *MedienPädagogik.* (5. Juli), 61–80. http://www.medienpaed.com/article/view/453. (Stand vom 06-06-2018).

Kant, Immanuel (1785): *Grundlegung zur Metaphysik der Sitten*, Werkausgabe Bd. VII, hrsg. v. Weischedel, Wilhelm, 2. Aufl., Frankfurt a. M.: Suhrkamp 1977.

Kant, Immanuel (1788): *Kritik der praktischen Vernunft*, Werkausgabe Bd. VII, hrsg. v. Weischedel, Wilhelm, 2. Aufl., Frankfurt a. M.: Suhrkamp 1977.

Kapp, Ernst (1877): *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten*. Braunschweig: Georg Westermann.

Kind, Sonja; Bovenschulte, Marc; Ehrenberg-Silies, Simone; Jetzke, Tobias & Weide, Sebastian (2017): *Social Bots. Thesenpapier zum öffentlichen Fachgespräch »Social Bots – Diskussion und Validierung von Zwischenergebnissen« am 26. Januar 2017 im Deutschen Bundestag*. o.O.: VDI/VDE Innovation + Technik GmbH.

Krafft, Peter M.; Macyy, Michael & Pentland, Alex “Sandy” (2016): *Bots as Virtual Confederates: Design and Ethics*. 2 Nov 2016, 1 (1–8) https://arxiv.org/pdf/1611.00447.pdf. (Stand vom 06-06-2018).

Leistert, Oliver (2017): Social Bots als algorithmische Piraten und als Boten einer techno-environmentalen Handlungskraft. In: Seyfert, Robert & Roberge, Jonathan (Hrsg.): *Algorithmuskulturen. Über die rechnerische Konstruktion der Wirklichkeit*. Bielefeld: transcript, S. 215–234.

Mill, John Stuart (1861): *Der Utilitarismus*, Stuttgart: Reclam 1985.

Murthy, Dhiraj; Powell, Alison B.; Tinati, Ramine; Anstead, Nick; Carr, Leslie; Halford, Susan J. & Weal, Mark (2016): Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital. *International Journal of Communication* 10, S. 4952–4971.

Schiller, Friedrich (1795): *Über die ästhetische Erziehung des Menschen. In einer Reihe von Briefen*. Stuttgart: Reclam 1979.

Turing, Alan M. (1950): Kann eine Maschine denken? In: Zimmerli, Walther Chr. & Wolf, Stefan (Hrsg.): *Künstliche Intelligenz. Philosophische Probleme*. Stuttgart: Reclam 1994, S. 39–78.

Können Chatbots lügen?

Claudia Paganini

Zusammenfassung

Chatbots systematisieren die natürliche menschliche Sprache nach vorgegebenen Regeln, harmonisieren und interpretieren Ungereimtheiten, gleichen Fehler aus, greifen schnell und effektiv auf umfangreiche Datenbestände zu, analysieren die Problemstellung über verschiedene Erkennungsmuster und geben schließlich eine Lösung aus. Sie informieren den Verbraucher über Produkteigenschaften, führen Dialoge und agieren als intelligente persönliche Assistenten. Aber können sie auch lügen? Um eine Antwort auf diese Frage zu finden, werden verschiedene Merkmale der Lüge herausgearbeitet: Sprache, Wahrheitsanspruch, Falschheit, Absicht zu täuschen, Fähigkeit zu täuschen, Schaden, Gegenüber und Kontext. Diese Kriterien werden in der Folge einzeln diskutiert und dahingehend hinterfragt, welche Eigenschaften ein Chatbot mitbringen müsste, damit man unter der jeweiligen Rücksicht sagen darf, dass er lügen kann. Dabei zeigt sich trotz einzelner Einschränkungen, dass es durchaus angemessen ist, von lügenden Chatbots zu sprechen.

Einleitung

Als Oliver Bendel und sein Team im Sommer 2016 auf die Frage nach dem Präsidenten der Vereinigten Staaten von ihrem Lügenbot die Antwort erhalten haben „Der Präsident der USA ist Donald Trump.“, war diese Aussage (noch) falsch und sie war lustig. Heute ist sie weder das eine noch das andere. Dieses Beispiel zeigt, dass die Zeit falsche Aussagen wahr machen und Heiteres zu bitterem Ernst werden lassen kann. Aber zeigt es auch, dass Chatbots lügen können?

Bendels Ansatz, Maschinenethik zu betreiben, indem er das, was zur Diskussion steht, konstruiert (Bendel et al. 2017), ist originell und – zumindest im Fall des Lügenbots – bis dato singulär. Chatbots, die – indem sie ihre Verknüpfung mit anderen Systemen nutzen und sich an Klassifikationen orientieren – wahre Aussagen durch falsche ersetzen können, gibt es allerdings auch in anderen Kontexten, so zum Beispiel im Bereich der Verkaufsschulung, im Coaching oder in der Psychotherapie. Dort wird gegenwärtig der Einsatz von Chatbots getestet, die sich in komplexen Gesprächssituationen bewähren müssen, indem sie Reaktionen vorhersehen, verworrene Informationen strukturieren oder eben auch lügen. Aber ist es tatsächlich so, dass diese Chatbots lügen, oder müsste man sich nicht vielmehr darauf beschränken festzustellen, dass sie verschiedene Techniken des Lügens beherrschen?

Dieser Frage werde ich im Folgenden nachgehen. Dabei werde ich mich nicht in erster Linie an unserem alltäglichen metaphorischen, stark anthropomorphisierenden Sprachgebrauch orientieren, sondern an dem, was die Philosophie an Definitionen anzubieten hat. Denn dass wir umgangssprachlich gerne davon reden, dass uns unsere Computer, Drucker

oder Handys ärgern wollen, bedeutet weder, dass wir unseren Artefakten Intentionalität zuschreiben wollten, noch bedeutet es, dass sie tatsächlich Intentionalität besitzen. Ein solches Reden bringt vielmehr zum Ausdruck, wie nah wir unseren Maschinen gerückt sind und zwar sowohl, was den konkreten Umgang mit ihnen, als auch, was die emotionale Haltung ihnen gegenüber betrifft.

Doch bleiben wir bei der Frage, ob Chatbots lügen können. Um eine befriedigende Antwort geben zu können, müssen zumindest die folgenden Schritte vollzogen werden. Wir müssen

1) klären, was eine Lüge ist;

2) aufzeigen, welche Voraussetzungen ein Mensch, ein Tier oder ein Artefakt mitbringen muss, damit er/es lügen kann;

3) diskutieren, ob Chatbots diese Voraussetzungen erfüllen.

Was dabei grundsätzlich keine Rolle spielen soll, sind moralphilosophische Überlegungen, d. h. Überlegungen dahingehend, ob es wünschenswert ist, dass Chatbots lügen, ob sie lügen dürfen oder gar sollen. Bei der Durchsicht der verschiedenen deskriptiven Zugänge zum Thema „Lüge" bzw. „lügen" wird sich aber zeigen, dass sich dieser normative Aspekt doch nicht ganz ausklammern lässt und zwar, dass er sich dort nicht ausklammern lässt, wo er Implikationen für die Definition hat, d.h. dass sich, je nachdem, von welchem moralphilosophischem Standpunkt man ausgeht, unterschiedliche Definitionen von „Lüge" ergeben.

In derartigen Fällen kommen moralphilosophische, normative Überlegungen quasi durch eine Hintertür ins Spiel, so zum Beispiel, wenn Augustinus in seiner Polarität zwischen Gut und Böse das Lügen als Form der Blasphemie verurteilt. Sofern das Lügen nämlich als moralphilosophisch bedenklich gilt, weil es eine Verletzung der Mensch-Gott-Beziehung impliziert, wird genau diese Verletzung der Mensch-Gott-Beziehung zu einem Merkmal von Lüge – wenngleich möglicherweise nicht zu dem entscheidenden. Je nachdem, welchen Stellenwert man diesem Merkmal nun aber zuschreiben wollte, würde die Antwort auf die Frage „Können Chatbots lügen?" mehr oder weniger deutlich negativ ausfallen. Denn gemeinhin nehmen wir an, dass Chatbots keine Gottesbeziehung haben. Damit aber könnten sie nicht lügen – zumindest nicht im vollen Sinn der Augustinus-Definition.

Glücklicherweise brauchen Schwierigkeiten, die sich aus der Position des Augustinus ergeben, uns nicht sonderlich zu beunruhigen. Denn erstens können seine (diesbezüglichen) Überlegungen nur für eine bestimmte Gruppe von Menschen gelten, nämlich für diejenigen, die an einen – wenn nicht christlichen, so zumindest monotheistisch gedachten – Gott glauben, und zweitens kann man nicht einmal innerhalb dieser Gruppe von einer breiten diesbezüglichen Zustimmung ausgehen. Fälle, wo normative Überlegungen die deskriptiven beeinflussen, gibt es aber dennoch und diese sollen im weiteren Textverlauf auch erwähnt werden.

Das unterschiedliche Verständnis von Lüge

Auf die Frage, was eine Lüge ist, kann man nicht nur inhaltlich unterschiedliche Antworten geben, man kann den Begriff auch enger oder weiter fassen. Beispiele für die eine wie für die andere Leseweise gibt es genug. In einem weiten Verständnis könnte man Lüge etwa als einen „kommunikativen Akt bezeichnen, der mit der Absicht zu täuschen unternommen wird“ oder als performative Kommunikation, „durch die neue emotive wie auch kognitive Wirklichkeiten erzeugt werden“ (Dietzsch 2010, S. 1467). Demgegenüber schon etwas präziser ist Sissela Boks Definition von einem „false statement made to another person with the intention to mislead“ (Kagan 1995, S. 515). Interessant ist dabei, dass Bok, obwohl sie sich mit der Mannigfaltigkeit des Lügens auseinandersetzt – sie spricht von „white lies“, „false excuses“, „inauthentic justifications“, „lies in crisis“, „lies to liars“, „lies to enemies“, „lies for the public good“, „lies to protect peers and clients“, „deceptive social science research“ und „lies to the sick and dying“ –, „Lüge“ letztendlich sehr ähnlich versteht wie das ursprüngliche lateinische *mendacium*. Nämlich: berechnend die Unwahrheit sagen. Deutlich offener war da noch das griechische *pseudos*, das neben der bewusst geäußerten Unwahrheit auch den Irrtum, die Fiktion oder die dichterische Ausschmückung im Blick hatte (Dietzsch 2010, S. 1468).

In Abhebung davon bezeichnen enge Definitionen einen Kommunikationsakt üblicherweise genau dann als Lüge, wenn 1) eine Behauptung p vorliegt, 2) der Sender p für falsch hält, 3) er die Absicht hat, den Empfänger p für wahr halten zu lassen, und 4) p tatsächlich falsch ist. Ob alle oder nur ein Teil dieser Bedingungen erfüllt sein müssen oder ob es sogar noch weitere Bedingungen braucht, ist Gegenstand der philosophischen Diskussion. Weitgehend unhinterfragt ist dagegen das Unternehmen, den Oberbegriff der Lüge weiter zu untergliedern, also verschiedene Arten von Lüge auszuweisen. Weit verbreitet sind in diesem Zusammenhang immer noch die Unterscheidungen von Thomas (Thomas von Aquin 1933ff., Summa Theologica II-II. Questio 110, art. IV, ad quint.) bzw. später Luther zwischen Scherzlüge, Nutzlüge und Schadenslügen (Luther 1518, S. 510-511), weniger bekannt dagegen sind Machiavellis strategische Täuschung (Machiavelli 1532, Kap. 18) oder die jesuitische *reservatio mentalis*, bei der ein geleistetes Versprechen in Gedanken uminterpretiert wird.

Für den Kontext dieses Beitrags relevant sind aber nicht so sehr die verschiedenen Arten des Lügens, sondern die Bedingungen einer Lüge, welche in den unterschiedlichen Definitionen enthalten sind. Sie bringen nämlich verschiedene Momente ins Spiel, welche ihrerseits Auswirkungen auf die Fragestellung „Können Chatbots lügen?“ haben. Deshalb erscheint es sinnvoll, im Folgenden diese Momente einzeln zu diskutieren und im Zuge dessen jeweils nach den erforderlichen Eigenschaften zu fragen, die Chatbots mitbringen müssten, damit sie den Bedingungen einer Lüge entsprechen können.

Sprache

Ein erstes umstrittenes Merkmal der Lüge ist die Sprachlichkeit. Während die einen nämlich nur dann von einer Lüge sprechen wollen, wenn eine explizit sprachliche Äußerung vorliegt, begnügen sich die anderen damit, dass es sich um eine einem zumindest „ähnlich ausgeprägtes Zeichensystem“ (Rott 2013, S. 125) zuordenbare Äußerung handeln muss. Wieder andere lassen – unter Verweis auf das Tierreich und das dort anzutreffende Täuschungsverhalten – die ganze Bandbreite der nonverbalen Kommunikation zu. Demnach würde es sich auch um eine Lüge handeln, wenn wir freundlich lächeln, obwohl wir eher Antipathie als Sympathie für unser Gegenüber hegen, beim Sex Erregung vortäuschen, obwohl wir uns langweilen etc.

Natürlich könnte man hier einwenden, dass man in solchen und ähnlichen Fällen eher von einer „Täuschung“ als von einer „Lüge“ sprechen sollte. Allerdings ist unklar, wie bzw. ob eine solche, hauptsächlich auf Konventionen – die sich ihrerseits wieder verändern können – gestützte Unterscheidung dazu beitragen könnte, das Phänomen Lüge besser zu begreifen. Für die hier diskutierte Thematik jedenfalls macht es keinen großen Unterschied, ob die Leitfrage mit dem Wörtchen „lügen“ oder „täuschen“ formuliert ist.

Doch zurück zur Sprachlichkeit, wo neben den bereits genannten Streitpunkten außerdem noch Uneinigkeit dahingehend herrscht, ob – so es sich um eine sprachliche Äußerung handelt – der gelogene Inhalt tatsächlich ausgesprochen werden muss. In sprachliche Gestalt gefasste Lügen können nämlich auch dadurch evoziert werden, dass mehrere wahre Aussagen auf eine Art und Weise verknüpft werden, die insgesamt dazu führt, dass der Empfänger einen falschen Sachverhalt für wahr hält, wie etwa in der Politik, wenn die Wählerschaft populistisch manipuliert wird. Ähnlich verhält es sich mit dem unkorrigierten Stehenlassen irrtümlicher Meinungen, dem Verschweigen wichtiger Inhalte, dem Weglassen relevanter Negativaspekte, dem Überhäufen des Adressaten mit richtigen, aber irrelevanten Informationen etc. Diese Überlegungen hängen de facto aber bereits eng mit dem nächsten Merkmal zusammen, dem Wahrheitsanspruch.

Für die Frage nach den lügenden Chatbots stellt das Kriterium der Sprache kein Problem dar. Denn Chatbots sind Software-Roboter, die gerade auf die Sprachbeherrschung hin entwickelt und programmiert sind. Mit Blick auf die Sprache fällt unsere Antwort also klar positiv aus. Eine nähere Auseinandersetzung mit diesem Kriterium könnte aber in Zukunft gewinnbringend sein, wenn Chatbots weiter humanisiert werden und sich nicht nur im Hinblick auf ihre Sprachbeherrschung, sondern auch in Mimik und Gestik stärker dem menschlichen Ausdrucksverhalten annähern.

Wahrheitsanspruch

Was weiters den Wahrheitsanspruch betrifft, gehen die Positionen sowohl hinsichtlich der Forderung auseinander, dass ein solcher gegeben sein soll, als auch hinsichtlich der angemessenen Form, in der ein Wahrheitsanspruch kommuniziert werden muss. Die engste

Interpretation besteht darin, einen in der Form eines Aussagesatzes explizit geäußerten Wahrheitsanspruch einzumahnen. Gerade in der Tradition der Sprachphilosophie wird aber zu Recht darauf hingewiesen, dass man sich, um eine sprachliche Äußerung adäquat zu erfassen, nicht primär von ihrer grammatikalischen Oberflächenstruktur leiten lassen, sondern die Aufmerksamkeit vielmehr auf die Intention des Sprechers richten sollte (Austin 1962). Behauptungen müssen nicht notgedrungen die Form eines Aussagesatzes annehmen, sie können auch als Frage oder Imperativ getarnt sein. Umgekehrt kann es sein, dass in Aussagesätze verpackte Äußerungen gerade keinen Wahrheitsanspruch transportieren, denken wir zum Beispiel an die Ironie, das Geschichtenerzählen, die Schauspielerei etc. Ob man all diese Fälle von unrichtigen Aussagen unter den Oberbegriff der Lüge subsummieren will, hat vor allem Auswirkungen auf die moralphilosophische Bewertung.

Mit Blick auf das Lügen von Chatbots aber ist der Wahrheitsanspruch als Kriterium weitgehend unproblematisch. Denn ein Chatbot, der lügen kann oder soll, wird sich vermutlich an anderen Systemen orientieren, die ihm für zuverlässig erachtete Informationen liefern können. Indem er aber mit diesen zuverlässigen Informationen arbeitet, entsteht in gewisser Weise ein Wahrheitsanspruch. Schwierigkeiten könnten hier einerseits auftreten, wenn man das Reden von einem Wahrheitsanspruch grundsätzlich problematisieren oder im Zuge der Auseinandersetzung mit Fragen der Realismus-Relativismus-Debatte überhaupt zurückweisen wollte. Letzteres erscheint aber nicht besonders zielführend zu sein, macht man sich bewusst, dass der hier zugrunde gelegte Wahrheitsbegriff nicht besonders stark ist. Vielmehr beschränkt er sich auf die weitgehend konsensfähige Annahme, dass es in einer Sprachgemeinschaften üblicherweise nicht völlig beliebig ist, was behauptet wird, sondern es vielmehr eine Reihe von Behauptungen gibt, von denen man annimmt, dass sie eher zutreffen als andere.

Andererseits entstehen Probleme mit dem Kriterium des Wahrheitsanspruchs, wenn ein Chatbot explizit dafür programmiert worden ist, die Unwahrheit zu sagen, bzw. dem Nutzer auch als eine solche Münchhausen-Maschine präsentiert wird. Dies lässt sich am besten über die Wirkung veranschaulichen, die beim Empfänger erzeugt wird. Ähnlich wie die Verstellung des Schauspielers, der in eine Rolle schlüpft und einen Charakter darstellt, der nicht der eigene ist, vom Publikum nicht als moralisches Ärgernis verurteilt, sondern als Ausdruck von Kunst interpretiert und als unterhaltsamer Zeitvertreib wertgeschätzt wird, hat Oliver Bendels Lügenbot wohl kaum einen Nutzer vor den Kopf gestoßen oder empört. Wer erwartet, angelogen zu werden, wer also nicht (mehr) mit einem Wahrheitsanspruch seitens des Senders rechnet, der reagiert auch nicht mit moralischer Entrüstung, sondern nimmt verstärkt andere, möglicherweise positive Aspekte der Unwahrheit wahr. Das ändert sich auch nicht, wenn man weiß, dass der Lügenbot zu 20% die Wahrheit sagt. Ob man in einem solchen Fall noch sinnvoll vom Lügen reden kann, ist daher m.E. fraglich.

Falschheit

Ein weiteres Kriterium, das auf den ersten Blick einleuchtet, bei näherem Hinsehen aber differenziert zu bewerten ist, stellt die Falschheit dar. Es ist zwar bereits gezeigt worden, dass nicht nur falsche, sondern eben auch wahre Aussagen zu falschen Überzeugungen führen können, die Frage, ob die evozierten Überzeugungen aber überhaupt falsch sein müssen, ist noch nicht gestellt worden. Es könnte nämlich auch sein, dass jemand in der falschen Meinung, dass p, einen anderen belügen will und also glauben lässt, dass $\neg$ p, es de facto aber der Fall ist, dass der Lügner sich täuscht, $\neg$ p zutrifft und der Belogene durch die Lüge dazu gebracht worden ist, einen wahren Sachverhalt für wahr zu halten. Analog könnte es passieren, dass ein Chatbot, der dafür programmiert worden ist, mithilfe von Klassifikationen falsche Aussagen zu produzieren, zufälligerweise auf eine richtige Aussage kommt, wenn er etwa in dem Satz „Pferde sind schnell." den Begriff „Pferd" nicht durch „Schnecke", sondern durch „Gepard" ersetzt – um ein allzu banales Beispiel zu bemühen.

Da es sich dabei aber um Sonderfälle, also um Ausnahmen von der Regel handelt und diese – durch das Sich-Täuschen des Sprechers hervorgerufenen – Ausnahmen Menschen wie Chatbots gleichermaßen betreffen können, sind sie für die hier zu diskutierende Fragestellung nicht weiter relevant.

Absicht zu täuschen

Anders verhält es sich dagegen bei der Täuschungsabsicht. Denn obwohl es auch in diesem Kontext Minderheitenpositionen gibt, wird üblicherweise davon ausgegangen, dass eine Lüge erst dann vorliegt, wenn der Sender die Absicht hat zu täuschen, sprich eine sogenannte *voluntas ad fallendum* besteht (Dietzsch 2010, S. 1468). Umgekehrt würde man nämlich gerade nicht von einer Lüge sprechen, wenn jemand eine falsche Aussage macht, diese selbst aber für richtig hält. Das Kriterium der Absicht ist im Zusammenhang mit einem Chatbot allerdings problematisch, wird die Frage nach der Intention doch gerne im Zusammenhang mit der Frage nach dem Bewusstsein gesehen. So ist man etwa geneigt, Tieren, die Täuschungsverhalten zeigen, auch Bewusstsein zu unterstellen. Geht man diesen Weg in die umgekehrte Richtung, würde aus der Annahme, dass Maschinen kein Bewusstsein haben, folgen, dass sie keine Absichten haben können, also auch keine Täuschungsabsicht. Wenn dem so wäre, müsste man feststellen, dass Chatbots nicht lügen können. Denn anders als bei manch einem der übrigen Kriterien gilt die Täuschungsabsicht in der Philosophie als eine notwendige (wenngleich nicht hinreichende) Bedingung für das Lügen.

Allerdings ist mit dem Kriterium der Absicht noch nicht gesagt, wer diese Absicht hat bzw. haben muss. Muss der Chatbot selbst die Absicht haben zu täuschen oder genügt es, dass es eine Person gibt, die diese Absicht hat bzw. hatte und dem Chatbot den entsprechenden Auftrag gibt? Denn ähnlich wie ein Saugroboter als mein Stellvertreter agiert, wenn er Staub saugt, und damit meine Intention, den Zustand meiner Wohnung zu verbes-

sern, realisiert, könnte es – um dem Kriterium der Absicht Genüge zu tun – ausreichen, wenn ein Chatbot jemandes anderen Absicht zu täuschen realisiert. Er wäre dann der Stellvertreter dieser Person, und es würde sich dabei keineswegs um etwas ganz Neues handeln, denn das Problem der Stellvertreter-Lüge ist in der Philosophiegeschichte immer wieder diskutiert worden, etwa wenn ein Vorgesetzter einen Untergebenen zwingt, für ihn die Unwahrheit zu sagen.

Wenn damit tatsächlich von einer Täuschungsabsicht die Rede sein kann, sollte zum Bewusstsein wohl auch noch der freie Wille hinzukommen. Dadurch aber wäre die Situation des Chatbots schon nicht mehr so weit von der seines menschlichen Konkurrenten entfernt. Betrachtet man nämlich die Situationen, in denen Menschen üblicherweise lügen, zeigt sich, dass nur ein kleiner Teil der Lügen wirklich aus freiem Willen geschieht. Viel häufiger sehen wir uns gezwungen zu lügen, um nicht unhöflich zu erscheinen, um größeren Schaden zu vermeiden, uns nicht zu blamieren etc.

Dazu kommt noch, dass die automatische Verknüpfung von Absicht und Bewusstsein keineswegs unproblematisch ist, wie die Materialismus-Debatte in der Philosophie des Geistes zeigt. Ohne dazu hier Stellung nehmen zu wollen, sei deshalb zumindest die Möglichkeit ins Auge gefasst, dass Chatbots – ohne dass wir ihnen ein Bewusstsein zusprechen müssen – im Lauf der Zeit und im Zuge der Erfahrungen, die sie machen, lernen, dass es für sie manchmal besser ist, die Unwahrheit zu sagen. Wenn aber ein auf sein eigenes Lernen hin konzipiertes autonomes System, das auf genetisch sich selbst programmierenden Algorithmen beruht, die „Wahl" treffen sollte zu lügen, wäre es dann wirklich noch zulässig, ihm jegliche Form von Absicht abzusprechen? Wenn – anders formuliert – kognitive Software-Agenten „durch intelligente Exploration und Verallgemeinerung ihre Umwelt zu verstehen lernen und die gelernten Modelle zur Handlungsplanung nutzen" (Decker 2016, S. 352), wird dann nicht das Beharren auf ihrem fehlenden Bewusstsein müßig?

Fähigkeit zu täuschen

Eng verknüpft mit der Absicht zu täuschen ist die Fähigkeit zu täuschen, welche in der Vergangenheit aufgrund der einseitigen Fixierung auf die negative moralphilosophische Bewertung der Lüge häufig zu wenig gesehen und schon gar nicht wertgeschätzt worden ist. Richtet man den Fokus aber auf diese Fähigkeit, mag die Lüge als „Zeichen sozialer Intelligenz" oder gar als „Triebfeder der Evolution" (Dietzsch 2010, S. 1467) gelten, und Johann Georg Hamann spricht davon, dass das Lügen zur Weisheit gehöre, die „irdisch, menschlich und teuflisch" sei (Hamann 1823, S. 143).

Lügen bedeutet nämlich nicht, einfach bloß $\neg$ p zu behaupten, wenn p der Fall ist, wie der Lügenbot mit seinen vielfältigen Lügenstrategien eindrucksvoll gezeigt hat. Lügen braucht Kreativität, exploratives Umgehen mit Wirklichkeit also und genau das ist es, worauf das Open-Ended-Machine-Learning der Gegenwart abzielt.

Schaden

Beim Aspekt des Schadens schließlich handelt es sich um ein Kriterium, das sich erst im Lauf der Zeit etabliert hat und zwar im Zuge einer stärker moralphilosophisch geprägten Auseinandersetzung mit der Lüge. Je rigoroser das theologische Verständnis wurde, je stärker man den Sündenaspekt unterstrich, desto mehr rückte auch das Kriterium „Schaden“ in den Fokus. Damit zeigt sich, wie bereits eingangs angedeutet, dass normative Zugänge bisweilen Rückwirkungen auf die deskriptive Ebene haben, ja diese verändern können. Besonders hervorgehoben wurde der Schaden übrigens von Immanuel Kant, der der Ansicht war, dass selbst eine Lüge, die keinen konkreten Schaden, sondern möglicherweise sogar einen Nutzen hervorruft, letztlich die ganze Menschheit schädige. Er schreibt: „[Die Lüge] schadet jederzeit einem Anderen, wenngleich nicht einem andern Menschen, doch der Menschheit überhaupt, indem sie die Rechtsquelle unbrauchbar macht“ (Kant 1883, S. 297).

In der gegenwärtigen Debatte hat sich u.a. die bereits genannte Sissela Bok mit dem langfristigen Schaden des Lügens auseinandergesetzt und zwar in Abhebung von einem möglicherweise ebenso gegebenen, jedoch kurzfristigen Nutzen. Sie nennt insbesondere den Verlust von Vertrauen auf der Seite der Belogenen bzw. von Selbst-Respekt und Realitätssinn auf der Seite des Lügners (Bok 1979). Übertragen auf die hier diskutierte Fragestellung jedenfalls lässt sich festhalten, dass Chatbots dem Menschen in Sachen Schaden um nichts nachstehen. Chatbots, die ihr Gegenüber belügen, erzeugen Schaden, indem nämlich Menschen getäuscht und möglichweise zu für sie nachteiligen Handlungen angeleitet werden, indem ihre Gefühle verletzt, ihr Vertrauen in Technik und digitale Kommunikation geschwächt wird usw. Ob es tatsächlich zu einem solchen Schaden kommen kann, hängt aber ganz wesentlich vom Gegenüber selbst ab.

Gegenüber

Im Zusammenhang mit diesem Gegenüber, dem Empfänger also, ist in der Vergangenheit nämlich nicht nur diskutiert worden, welche Hörer überhaupt das Recht auf die Wahrheit haben (HR 2013, S. 125), sondern auch, was für Voraussetzungen sie mitbringen müssen, damit sie überhaupt belogen – und insofern möglicherweise geschädigt – werden können. Letztere Überlegung wird klarer, wenn man sich einmal mehr bewusst macht, dass wir es im Alltag mit einer ganzen Reihe von Sprachspielen zu tun haben, in denen nicht die Wahrheit gesagt wird, die aber dennoch aus einer moralphilosophischen Perspektive heraus als absolut unbedenklich gelten können: Ironie, Scherz, Kabarett, Roman, Rollenspiel, Schauspiel etc.

Wie schon im Zusammenhang mit dem Wahrheitsanspruch angedeutet, muss der Adressat einer Lüge davon ausgehen, dass der Lügner die Wahrheit sagt, damit überhaupt eine Lüge zustande kommen muss. Natürlich findet man auch hier Denker, die diese Position nicht teilen, die genannten Beispiele zeigen m.E. jedoch eindrucksvoll ihre Plausibilität. Insofern müsste man schlussfolgern, dass ein Chatbot wie der Lügenbot, von dem der Emp-

fänger die Unwahrheit erwartet, eigentlich nicht lügen kann und deshalb auch keine unmoralische Maschine ist. Sehr wohl aber lügen Chatbots (mit Rücksicht auf die Erwartungshaltung ihres Gegenübers), wenn sie in einer Situation, wo eine wahre Aussage – wie etwa der Wetterbericht für die Tourismusregion Tirol – erwartet wird, eine falsche Aussage liefern, so zum Beispiel die um ein paar Sonnenstunden und mit einer niedrigeren Niederschlagswahrscheinlichkeit geschönte Vorschau. Auch lügen Chatbots, wenn sie sich in Verhandlungssituationen mit ihren menschlichen Gesprächspartnern im Taktieren messen sollen, wenn sie also in einen bestimmten Kontext gebracht werden, in dem offiziell zwar Ehrlichkeit eingefordert wird, man mit dieser Ehrlichkeit allein aber nur schwerlich Erfolg haben kann.

Kontext

Was diesen gesellschaftlichen Kontext betrifft, so ist er spätestens seit Nietzsches Kritik – Wahrhaftigkeit bestünde in der „Verpflichtung, nach einer festen Convention zu lügen, schaarenweise in einem für alle verbindlichen Sinn zu lügen" (Nietzsche 1973, S. 375) – Gegenstand der Diskussion. In eine ähnliche Richtung wie Nietzsche geht beispielsweise Max Scheler, der in diesem Zusammenhang meint: „Wer (organisch) verlogen ist, braucht nicht mehr zu lügen" (Scheler 1955, S. 67). Auch wenn diese Positionen möglicherweise als überspitzt einzustufen sind, zeigen sie doch, dass eine individualethische Fokussierung auf den Lügner naiv bleibt. Denn der Einzelne agiert immer als Teil eines Kollektivs, und es gilt nicht zuletzt dieses Kollektiv zu analysieren, wenn man das Phänomen der Lüge besser begreifen will.

In Abhebung zu Nietzsche und Scheler lässt sich aber feststellen, dass eine Sprachgemeinschaft grundsätzlich auf ein (zumindest) gewisses Maß an Wahrhaftigkeit und Vertrauen ausgerichtet sein muss, damit sie als solche erhalten bleiben kann. Die Sprache S einer Population P würde nämlich jegliche Funktionalität verlieren, wenn ihre Konventionen an Wahrhaftigkeit und Vertrauen (gänzlich) aufgegeben würden, ja diese werden gerade durch das Interesse der Population an einer effektiven Kommunikation aufrecht erhalten (Lewis 1975, S. 169). Nichtsdestotrotz gibt es eine ganze Reihe von Kontexten, in denen ritualisiert gelogen wird, so etwa auf die Frage, wie es einem gehe, wenn ein Produkt beworben oder verkauft werden soll, bei Empfehlungsschreiben für entlassene MitarbeiterInnen etc. Auch sind bestimmte Lebensbereiche als besonders verletzlich anzusehen, d.h. in ihnen wird besonders häufig gelogen – wie z.B. in der Politik, Medizin oder Wirtschaft, wo massive Geldinteressen dahingehend bestehen, bestimmte Sachverhalte als wahr erscheinen zu lassen.

Für die Frage, ob Chatbots lügen können, sind diese Beobachtungen auf zweierlei Weise relevant. Erstens, weil man festhalten kann, dass ein Chatbot streng genommen nur in einem gesellschaftlichen Umfeld lügen kann, in dem zumindest gewisse Standards der Wahrhaftigkeit und Ehrlichkeit gelten. Von daher ist die Angst vor einem inflationären Lügen von Chatbots eindeutig zu relativieren. In einer solchen Sprachgemeinschaft wür-

den die Nutzer ihren Chatbots ganz einfach nicht mehr glauben. Dass diese Sprachgemeinschaft zugleich völlig dysfunktional wäre, ist ein anderes Problem, das hier nicht diskutiert werden soll. Zweitens zeigt sich, dass gerade die Auseinandersetzung mit der „künstlichen“ Lüge dazu beitragen kann, die Aufmerksamkeit für ritualisiertes und formalisiertes Lügen sowie für sensible Lebensbereiche zu schärfen und in eine konstruktive Auseinandersetzung mit den eigenen Vorstellungen zu treten, wie ein gesellschaftlicher Hintergrund beschaffen sein muss, damit der Einzelne – und eben auch ein Chatbot – vor diesem Hintergrund bisweilen lügen kann.

Fazit

Je nachdem, für welche Definition man sich entscheidet, d. h. welche Kriterien man für relevant erachtet, wird man die typisch menschlichen Eigenschaften, die für das Lügen benötigt werden, beim Chatbot finden oder eben nicht. Bei Beachtung der hier vorgestellten wesentlichen Merkmale jedoch scheint die Antwort auf ein klares Ja hinauszulaufen. Wer daran zweifelt, weil er meint, dass eben doch bestimmte darüber hinaus gehende Eigenschaften nicht vorhanden sind, sollte bedenken, dass Chatbots zwar auf der einen Seite manche menschliche Fähigkeiten nicht besitzen, auf der anderen Seite aber über Fähigkeiten verfügen, die wir Menschen entbehren, die für das Lügen aber sehr nützlich sein können wie zum Beispiel die Parallelisierbarkeit.

Chatbots können lügen und wenn sie uns ähnliche Partner sein sollen, dann müssen sie auch lügen. Denn die menschliche Kommunikation ist um einiges vielschichtiger und schillernder, als ein bloßes Abbild für wahr gehaltener Sachverhalte zu sein. Mit lügenden Chatbots aber adäquat umgehen zu können, erfordert Medienkompetenz, so wie es analog im nicht-digitalen Alltag der Menschenkenntnis bedarf, um richtig einschätzen zu können, wann jemand mir gegenüber aufrichtig ist, wann er mit mir scherzt oder mich täuschen will. Dieser Befund ist nicht beunruhigend, zeigt jedoch, dass nicht nur die Chatbots Zeit brauchen, um ihre Kommunikation mit dem Menschen zu perfektionieren, sondern dass auch wir Zeit und Gelegenheiten brauchen, Verhaltensunsicherheiten zu überwinden.

Literatur

Augustinus (2013): De mendacio. Contra mendacium. Contra Priscillianistas. In: Brachtendorf, Johannes & Drecoll, Volker Henning (Hrsg.): *Opera-Werke*. Band 50. Paderborn: Schöningh.

Austin, John Langshaw (1962): *How to do things with words*. Cambridge: Harvard University Press.

Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2017): Towards Kant Machines. In: *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, S. 7–11.

Bok, Sissela (1979): *Lying: Moral Choice in Public and Private Life*. New York: Vintage Books.

Decker, Michael (2016): Roboterethik. In: Heesen, Jessica (Hrsg.): *Handbuch Medien- und Informationsethik*. Stuttgart: Metzler, S. 351-357.

Dietzsch, Steffen (2010): Lüge. In: Sandkühler, Hans Jörg (Hrsg.): *Enzyklopädie Philosophie*. Band 2. I-P. Hamburg: Felix Meiner Verlag, S. 1467-1470.

Hamann, Johann Georg (1823): *Schriften*. Band IV. Berlin: G. Reimer.

Kagan, Richard (1995): Lying. In: Roth, John K. (Hrsg): *International Encyclopedia of Ethics*. London: Fitzroy Dearborn Publishers, S. 514-515.

Kant, Immanuel (1838): Über ein vermeintes Recht aus Menschenliebe zu lügen. In: Rosenkranz, Karl & Schubert, Friedrich Willhelm (Hrsg): *Sämmtliche Werke.* Siebenten Theils. Leipzig: Leopold Voss.

Machiavelli, Niccolò (1532): *Il principe*. Florenz.

Nietzsche, Friedrich (1973): Über Wahrheit und Lüge im außermoralischen Sinn. In: Colli, Giorgio & Montinari, Mazzino (Hrsg.): *Werke*. Kritische Gesamtausgabe III/2. Berlin, New York: Walter de Gruyter.

Luther Martin (1883): Praeceptum octavum. In: *Decem praecepta Wittenbergensi praedicata populo*, 1518, Werke. Krit. Gesamtausgabe. I. Weimar.

Thomas von Aquin (1933ff): *Die deutsche Thomas-Ausgabe* (Summa theologica), Übers. von Dominikanern u. Benediktinern Deutschlands u. Österreichs. Graz: Styria.

Rott, Hans (2013): Lüge. In: Mittelstraß, Jürgen (Hrsg.): *Enzyklopädie Philosophie und Wissenschaftstheorie*. Band 5: Log-N. Stuttgart, Weimar: J.B. Metzler.

Scheler, Max (1955): *Vom Umsturz der Werte*. Bern, München: Francke Verlag.

Kurzbiografien der Autorinnen und Autoren

Oliver Bendel, Prof. Dr., ist studierter Philosoph, studierter Informationswissenschaftler und promovierter Wirtschaftsinformatiker. Er lehrt und forscht an der Hochschule für Wirtschaft FHNW mit den Schwerpunkten Wissensmanagement, Wirtschafts-, Informations- und Maschinenethik.

Thorsten Büchner war von Geburt an stark sehbehindert und ist während seiner Schulzeit an der Deutschen Blindenstudienanstalt e.V. (blista) erblindet. Er arbeitet heute im Bereich Öffentlichkeitsarbeit der blista und gestaltet dort u. a. das akustische Infomagazin der Deutschen Blinden Bibliothek, den „Kopfhörer". Er ist zugleich Mitglied des Marburger Stadtparlaments, Mitglied im Behindertenbeirat der Universitätsstadt Marburg und Sprecher des Fahrgastbeirats der Universitätsstadt Marburg und des Landkreises Marburg Biedenkopf.

Theo Hug, Dr. phil., Professor für Erziehungswissenschaft am Institut für Medien, Gesellschaft und Kommunikation der Universität Innsbruck mit Schwerpunkt Medienpädagogik und Kommunikationskultur, Sprecher des interfakultären Forums *Innsbruck Media Studies* an der Universität Innsbruck.

Timo Kaerlein, Dr., ist wissenschaftlicher Koordinator am Graduiertenkolleg Locating Media der Universität Siegen. Aktuelle Forschungsschwerpunkte: Digitale Nahkörpertechnologien, Interfaces: Theorie, Ästhetik und Geschichte, Medienkulturen der Obsoleszenz, Social Robotics.

Andrea Knaut, Dr., Diplom-Informatikerin, hat über Fehler biometrischer Fingerabdruckerkennungssysteme an der Humboldt-Universität zu Berlin promoviert. Sie lehrt dort und an der HTW Berlin seit einigen Jahren über informatische Bildung und Kritische Informatik und ist aktives Mitglied der Arbeitsgruppe „Internet und Gesellschaft" der Gesellschaft für Informatik (GI) und des Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FIfF).

Rainer Leschke, Prof. Dr., Medienwissenschaftler an der Philosophischen Fakultät der Universität Siegen. Seine Arbeitsschwerpunkte liegen insbesondere im Bereich der Medientheorie und der Medienphilosophie. Weblink: http://www.rainerleschke.de/

Netaya Lotze, Dr., ist Sprachwissenschaftlerin mit Schwerpunkt „Neue Medien" und Akademische Rätin an der Westfälischen Wilhelms-Universität Münster, Germanistisches Institut, Abteilung Sprachwissenschaft; aktuelle Arbeitsschwerpunkte: Mensch-Maschine-Interaktion, Sprache und Neue Medien, Jugendsprache, Soziolinguistik, Onomastik.

Tilmann Märk, Univ.-Prof. Dr. Dr. h.c. mult., Rektor der Universität Innsbruck.

Claudia Paganini, Dr. phil., Institut für Christliche Philosophie der Universität Innsbruck; aktuelle Arbeitsschwerpunkte: Medienethik, Technikethik, Tierethik, Medizinethik.

Günther Pallaver, Dr. jur., Dr. phil., Professor für Politikwissenschaft mit dem Schwerpunkt Medien und politische Kommunikation am Institut für Politikwissenschaft sowie Leiter des Instituts für Medien, Gesellschaft und Kommunikation an der Universität Innsbruck.

Hermann Petz, Mag., Vorstandsvorsitzender der Moser Holding AG, Innsbruck.

Clemens Pig, Dr., Vorsitzender der Geschäftsführung, geschäftsführender Vorstand der APA – Austria Presse Agentur, Wien.

Imke Troltenier, Dr., hat sich im Bereich der Verhaltensforschung an der Philipps-Universität Marburg promoviert, arbeitete viele Jahre in der Entwicklung und Beratung Europäischer Bildungsprojekte für benachteiligte Gruppen und leitet heute die Öffentlichkeitsarbeit der Deutschen Blindenstudienanstalt e.V. (blista), einem Bundesweiten Kompetenzzentrum für Menschen mit Blindheit und Sehbehinderung.

Immer öfter kommuniziert Software mit und für uns. Bots übernehmen oft unbemerkt komplexe Aufgaben und virtuelle Assistenzen verarbeiten die gesprochene Sprache. Die Technologieentwicklung gestattet uns mittlerweile, die Dienste von digitalen, oftmals sprachbasierten Assistenzsystemen in unseren Alltag zu integrieren. Besonders im Bereich der politischen Propaganda und Public Relations sind Social Bots seit einiger Zeit zu zweifelhafter Berühmtheit gelangt. Eingesetzt als Instrument zur Meinungsmache verbreiten sie automatisiert Nachrichten in sogenannten „sozialen" Netzwerken. Der interdisziplinär ausgerichtete Band befasst sich mit dieser Thematik und sucht nach Antworten auf Fragen wie die folgenden: Wie können wir die Interaktion mit künstlichen Intelligenzen von jener mit Menschen unterscheiden? Inwieweit ist das wichtig? Was bedeutet diese Entwicklung für das Alltagsleben und für den Bereich der Kommunikation? Welche Bedeutung hat der verbreitete Einsatz von Bots als neue Schnittstelle zwischen Mensch und Maschine?


ISBN 978-3-903187-29-0
9 783903 187290